International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2025: 7.089

Exemplification of Local Regression at Leaf Nodes of Decision Tree

A. D. Mankar[1], S. D. Bhoite[2], K. G. Kharade[3], K. A. Raskar[4]

[1]Department of Computer Science, Tuljaram Chaturchand College of Arts, Science and Commerce, Baramati, Maharashtra, India
Email: *abhijeetmankar[at]gmail.com*

[2]School of Computer Science and Applications, Chhatrapati Shahu Institute of Business Education and Research, Kolhapur, Maharashtra, India
Email: *sdbhoite05[at]gmail.com*

[3]Department of Computer Science, Shivaji Univesity, Kolhapur, Maharashtra, India
Email: *kgk_csd[at]unishivaji.ac.in*

[4]Department of Science and Computer Science, MIT College(Alandi), Pune, Maharashtra, India
Email: *kirtiraskar1711[at]gmail.com*

Abstract: ***CART algorithm is used to generate decision trees. CART algorithm is implemented in RStudio by means of RPART package. In a regression tree created using RPART, the average value of the dependent variable column for the observations located in each terminal node is shown. There may exist different patterns at different leaf nodes. This paper proposes to apply linear regression to the leaf nodes of the regression tree developed by RPART. The MSE (mean squared error) of all leaf nodes in regression tree obtained by RPART is higher as compared to the MSE (mean squared error) of applying local regression to RPART leaf nodes.***

Keywords: Decision trees, Regression, Leaf nodes, Mean squared error, CART, rpart

1. Introduction

Many authors have previously proposed to generate models in the leaf nodes of the tree instead of predicting the mean. Notable tree algorithms proposed are SECRET, GUIDE, SUPPORT [4]. Torgo developed an inductive system (HTL) capable of learning regression trees and is able to utilize any of the nearest neighbor approximation, linear models, kernel regression models in the tree leaves [2]. Loh reviewed available algorithms for classification and regression trees. The author compared GUIDE, CART and M5 regression tree algorithms in the research paper [1]. Chaudhuri et al. introduced new method called SUPPORT for tree structured regression and matched its prediction mean square error (PMSE) with those of other methods. CART selects its splits on the basis of the degree of reduction in residual sum of squares whereas SUPPORT picks its splits by analyzing the distributions of the residuals [3]. Quinlan described M5, a system to learn models capable of predicting values. Similar to CART, M5 constructs tree-based models. CART employs regression trees that contain values in their leaf nodes. In case of M5, the trees built can have multivariate linear models. As per the author, M5 produces model trees that are generally considerable smaller as compared to regression trees produced by CART [5]. Malerba et al. presented Stepwise Model Tree Induction (SMOTI). The major aspect of SMOTI is induction of trees accompanied by two kinds of nodes namely regression nodes and split nodes. Regression nodes are responsible for carrying out straight-line regression exclusively, while split nodes are tasked with partitioning the sample space. There is a weight linked to type of node which allows user to choose either local regression or global regression [6]. Gadekar and Gore proposed a general form of regression model. The major aspect of the proposed model is linearity in coefficients and non-linearity of predictors in the form of link functions [7]. Czajkowski, M., & Kretowski, M. stated the difficulty in choosing the type of decision tree for everyday problems. According to the authors, the superior representation cannot be chosen beforehand. The authors performed experiments on synthetic and real-life datasets [8]. Dobra and Gehrke proposed for each intermediate node to apply the EM algorithm for Gaussian mixtures in order to find two clusters in data. They introduced SECRET (Scalable EM and Classification based Regression Trees) algorithm involving linear models in the leaves which needed very less computing power on large datasets. The authors compared their proposed algorithm with GUIDE algorithm in terms of accuracy and stated that the proposed algorithm is as accurate as GUIDE if normal splits are used [9].

CART is suited to both classification and regression problems [10]. In case of regression problem, the leaf nodes contain number of observations from the dataset. However, in order to represent the expected value for that leaf node, each leaf node displays the mean of the response variable column for all of the observations that were located there. This average(mean) operation for the leaf nodes in regression tree in CART algorithm gives rise to the idea to apply linear regression to the leaf nodes of CART tree.

To implement the above idea, one has to get access of all leaf nodes created in the CART decision tree through RPART package. The leaf nodes generated by RPART package can be accessed by ‘which’ and ‘where’ clause in RStudio. For experimentation purpose, five datasets from UCI Machine Learning Repository are used in this study. The mean squared error using RPART and mean squared error by applying local regression to the leaf nodes obtained using RPART is compared for all five datasets. By

observing the barplot of all five datasets, the MSE of RPART is comparatively more as compared to the MSE of applying local regression to RPART leaf nodes.

2. Data

A study was conducted in order to compare the results produced by rpart model in RStudio with the results produced by individually applying linear regression to each leaf nodes in rpart tree. Data were produced and analyzed using the RStudio 2023.06.0+421. For conducting experiments, five different datasets have been considered in this study. The datasets for this study were obtained from the UCI Machine Learning Repository. The datasets used are of different size, small, medium, large in terms of number of observations. The smallest dataset used is 'Concrete Slump Test' with 103 observations while the largest dataset used is 'Combined Cycle Power Plant' dataset with 9568 observations. The datasets exercised are expressed in Table 1.

Table 1: Overview of the datasets utilized in the research

Name of the dataset	Subject Area (as per UCI ML Repository)	Target variable considered	Total no. of observations
Iris ('Species' variable removed) [11]	Life Science	Petal. Width	150
Liver Disorders [12]	Health and Medicine	Drinks	345
Concrete Slump Test ('No', 'FLOW (cm)', 'Compressive Strength (28-day) (Mpa)' variables removed) [14]	Computer Science	SLUMP	103
QSAR Fish Toxicity [13]	Physics and Chemistry	LC50	908
Combined Cycle Power Plant [15]	Computer Science	PE (net hourly electrical energy output)	9568

3. Methodology

RStudio's RPART package implements the CART algorithm. When rpart is used to build regression tree, it recursively partitions the dataset. At the leaf nodes, the prediction is shown as the average value of the response variable column for all the observations found in the leaf node. We calculated mean squared error (MSE) for all the leaf nodes found in rpart tree. There may exist different patterns at different leaf nodes. For that reason, we then applied linear regression to all the observations in each leaf nodes. After this the mean squared error (MSE) is then calculated. By comparing the two MSEs, we observe that MSE of rpart is more than the MSE of applying local regression to each leaf nodes obtained through RPART. The 'Iris' dataset contains 'Sepal.Length', 'Sepal.Width', 'Petal.Length', 'Petal.Width' and 'Species' variables. We focus on numeric variables, hence 'Species' variable is removed from the dataframe before we apply rpart to it. In this dataset, we regard 'Petal.Width' as the dependent variable. For 'Concrete Slump Test' dataset, there are three response(target) variables namely, 'SLUMP', 'FLOW(cm)' and 'Compressive Strength(28-day)(Mpa)'. We considered 'SLUMP' response variable and hence remaining two are removed from the dataset. Also, 'No' is variable which has serial number of observations stored, is removed from the dataset. The steps in finding the MSE using RPART and MSE using local regression are mentioned below:

1. generate rpart decision tree
2. store all the leaf nodes of the tree in an array variable, i.e. leaves
3. initialize loop counter, i.e. i to 1
4. while i <= length(leaves)
   begin
   4.1 build dataframe for i^{th} leaf node containing all observations present in the i^{th} leaf node
   4.2 apply linear regression to i^{th} leaf node by considering the dataframe of the i^{th} leaf node
   4.3 increment i by 1
   end
5. initialize agmse_rpart and agmse_rplr to zero
6. initialize loop count, i.e. i to 1
7. while i<= length(leaves)
   begin
   7.1 initialize sum_rpart, sum_rplr to zero
   7.2 find the observed values in the ith leaf node
   7.3 compute the predicted_rpart value for the ith leaf node
   7.4 compute the predicted_rplr value for the ith leaf node
   7.5 compute the diff_rpart=observed-predicted_rpart
   7.6 compute the diff_rplr=observed-predicted_rplr
   7.7 compute the sum_rpart=sum_rpart+(diff_rpart^2)
   7.8 compute the sum_rplr=sum_rplr+(diff_rplr^2)
   7.9 compute the mse_rpart=sum(sum_rpart)/Number of rows in ith leaf node
   7.10 compute the mse_rplr=sum(sum_rplr)/Number of rows in the ith leaf node
   7.11 agmse_rpart=agmse_rpart+mse_rpart
   7.12 agmse_rplr=agmse_rplr+mse_rplr
   7.13 display mse_rpart
   7.14 display mse_rplr
   7.15 increment i by 1
   end
8. display agmse_rpart
9. display agmse_rplr

4. Results

We computed mean squared error (MSE) for RPART and RPART-LR(recursive partitioning local regression). The MSEs obtained for all datasets are described in Tables 2 to 6. For all five datasets chosen, we found that MSE using RPART-LR is less as compared to its counterpart. Further the decision trees obtained by applying RPART to respective datasets are shown in Fig. 1 to 5. The leaf nodes in the decision trees show the number of observations present in the leaf node along with the predicted value(mean) for the response variable.

Table 2: MSE (using rpart) Vs MSE (using rpart-lr) (Iris dataset)

Dataset	Iris		
Leaf node number	No. of observations	MSE (using rpart)	MSE (using rpart-lr)
2	50	0.010884	0.009241281
4	45	0.0304	0.01109387
6	13	0.009353846	0.003198035
7	42	0.06438095	0.0489532
Total	150	0.1150188	0.07248639

Table 3: MSE (using rpart) Vs MSE (using rpart-lr) (Liver Disorders dataset)

Dataset	Liver Disorders		
Leaf node number	No. of observations	MSE (using rpart)	MSE (using rpart-lr)
4	95	3.543657	3.225233
6	48	2.65	2.49569
8	9	0.3052632	0.138538
9	13	1.136842	0.9341181
12	60	3.120132	3.058595
13	35	3.803008	3.381216
15	30	3.615439	3.281152
16	12	0.7254386	0.3378033
19	8	0.2628289	0.1101526
20	16	1.839309	1.349097
21	19	5.810803	3.957396
Total	345	26.81272	22.26899

Table 4: MSE (using rpart) Vs MSE (using rpart-lr) (Concrete Slump Test dataset)

Dataset	Concrete Slump Test		
Leaf node number	No. of observations	MSE (using rpart)	MSE (using rpart-lr)
3	17	52.16955	12.68298
4	12	27.15319	6.859806
6	8	38.375	0
8	17	17.2128	6.832383
10	17	56.20934	22.79263
11	32	4.762753	2.199148
Total	103	195.8826	51.36694

Table 5: MSE (using rpart) Vs MSE (using rpart-lr) (QSAR Fish Toxicity dataset)

Dataset	QSAR Fish Toxicity		
Leaf node number	No. of observations	MSE (using rpart)	MSE (using rpart-lr)
6	68	0.7595504	0.5547086
7	26	0.3202429	0.1142128
8	40	0.7331515	0.533277
12	102	1.54771	1.439641
13	38	0.8667647	0.5661028
14	37	0.4933958	0.3548283
15	7	0.07476005	0.00003736184
17	183	1.19463	0.8948146
18	97	1.387422	1.049406
21	141	1.148371	0.960314
22	28	0.5705296	0.3787277
24	84	1.397224	1.2163
26	46	0.6446699	0.4789298
27	11	0.1168873	0.0188108
Total	908	11.25531	8.56011

Table 6: MSE (using rpart) Vs MSE (using rpart-lr) (Combined Cycle Power Plant dataset)

Dataset	Combined Cycle Power Plant		
Leaf node number	No. of observations	MSE (using rpart)	MSE (using rpart-lr)
4	2247	22.38926	16.12437
5	1593	18.26722	11.25857
6	1791	33.73466	11.26336
8	2200	36.32925	16.51229
9	1737	32.80046	17.92434
Total	9568	143.5208	73.08292

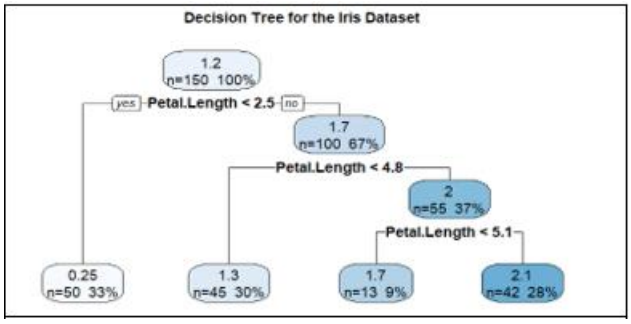

Figure 1: Decision tree for Iris dataset

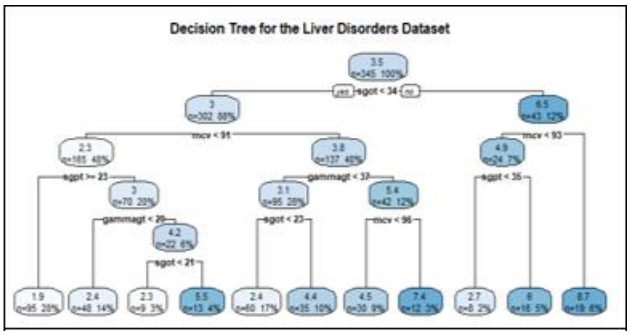

Figure 2: Decision tree for Liver Disorders dataset

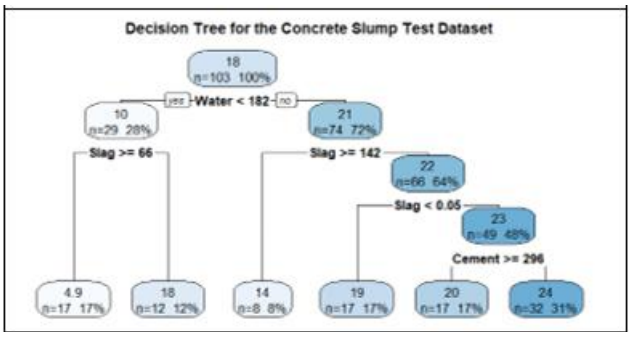

Figure 3: Decision tree for Concrete Slump Test dataset

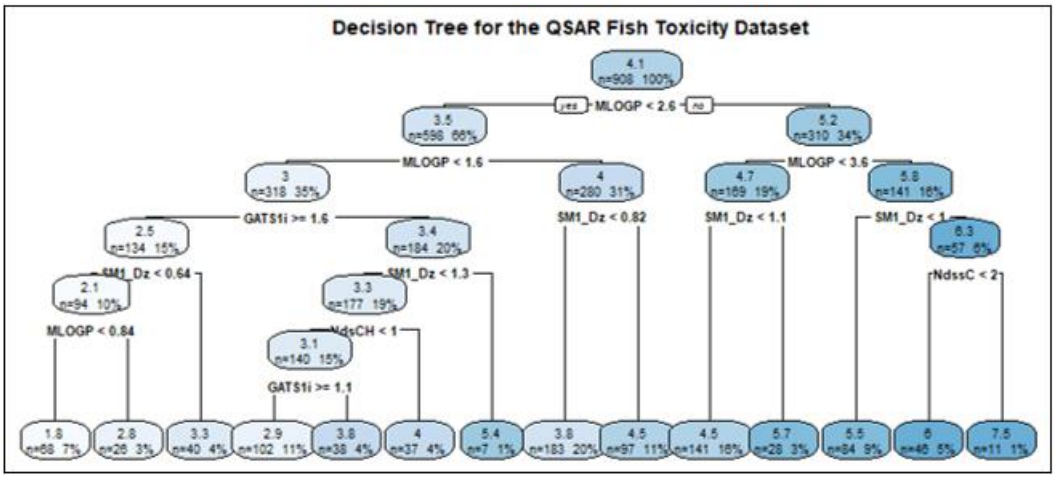

Figure 4: Decision tree for QSAR Fish Toxicity dataset

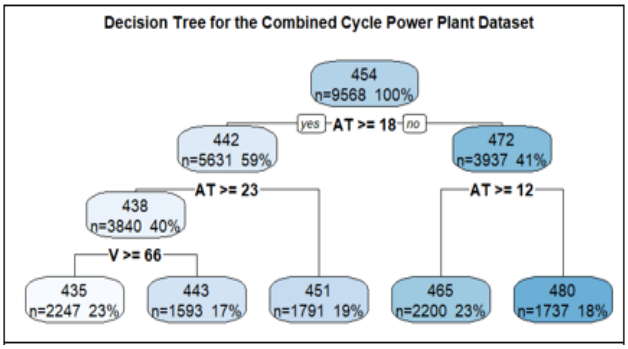

Figure 5: Decision tree for Combined Cycle Power Plant dataset

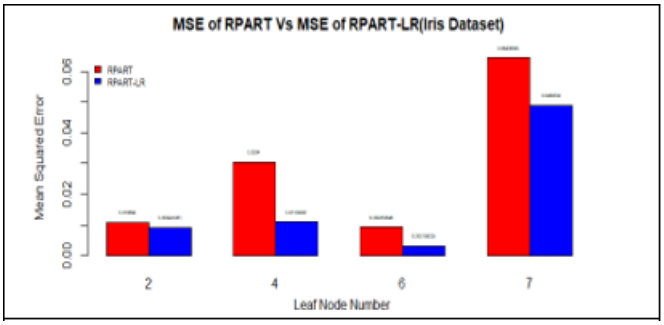

Figure 6: MSE of RPART Vs MSE of RPART-LR(Iris)

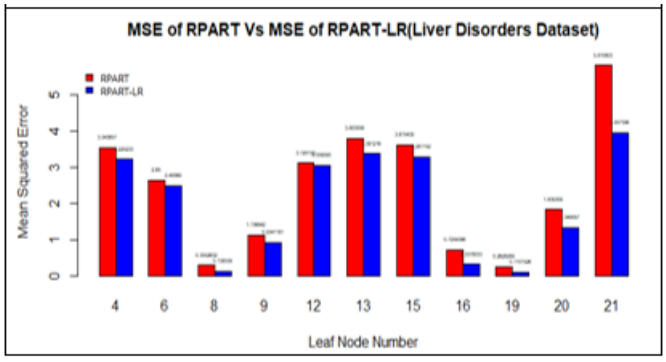

Figure 7: MSE of RPART Vs MSE of RPART-LR (Liver Disorders)

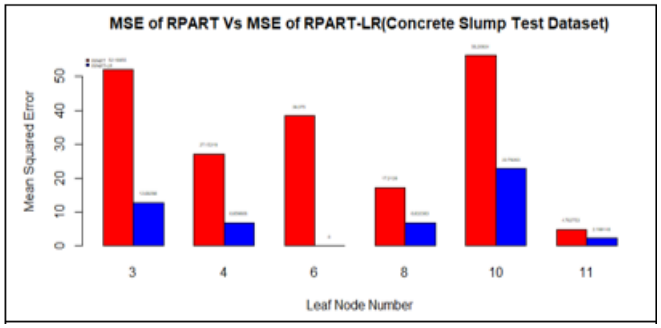

Figure 8: MSE of RPART Vs MSE of RPART-LR (Concrete Slump Test)

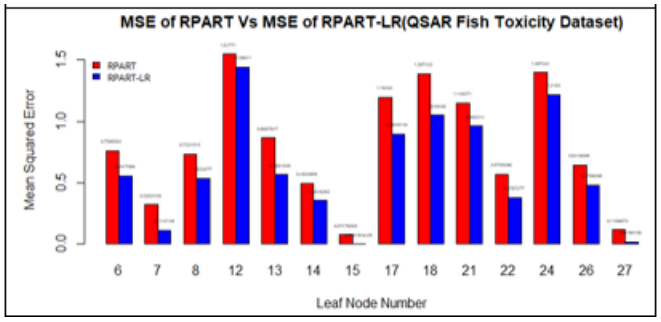

Figure 9: MSE of RPART Vs MSE of RPART-LR (QSAR Fish Toxicity)

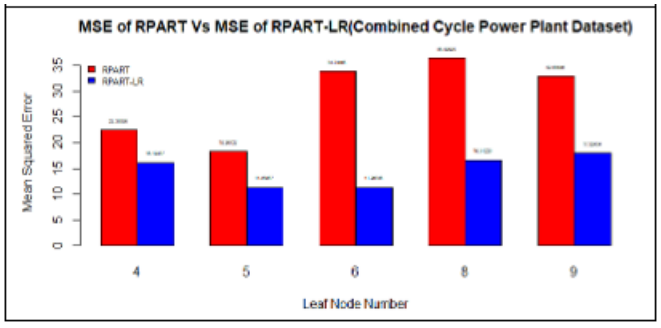

Figure 10: MSE of RPART Vs MSE of RPART-LR (Combined Cycle Power Plant)

We find that, from Fig. 6 to 10, the MSE produced by applying rpart to the dataset is consistently higher than the MSE obtained by applying rpart-lr(local regression) to the dataset, though the datasets utilized in this study are of varied sizes.

5. Conclusion

As observed from Tables 2,3,4,5,6 the MSE of RPART is more as compared to applying local regression to leaf nodes obtained through RPART. We call method of applying local regression to leaf nodes of RPART as RPART-LR in this paper. While developing a regression tree, the rpart package in RStudio recursively partitions the dataset, but at the leaf nodes the prediction is taken as the average of the response variable for all observation discovered within the corresponding leaf node. The regression tree obtained after applying rpart treats all observations in each leaf nodes as homogeneous. There may exist different patterns at different leaf nodes. Hence, we propose a new method for developing predictive models at the leaf level nodes. The proposed method will first start from the entire dataset (all observations) as the root node of the decision tree. The proposed method will then find the choice of splitting variable for the root node and the split point. The process of recursive partitioning is repeated till the size of the leaf node is reached to predefined value. Then for each leaf node of the decision tree local regression will be applied to obtain predictions for the individual observations found at the leaf nodes. Finally MSE (mean squared error) and aggregate MSE of all leaf nodes will be computed.

6. Future Scope

In this paper, we have used linear regression as local regression at leaf nodes. But different non-linear regressions can also be used as local regression at leaf nodes.

Funding
No funding was received for conducting this study.

Data availability
The data used in this study were downloaded from [11-15].

Statements and Declarations

Competing Interests
The authors here declare that they have no relevant financial or non-financial interests to disclose.

References

[1] Loh, W. Y. (2011) Classification and regression trees. Wiley interdisciplinary reviews: data mining and knowledge discovery 1(1):14-23.
[2] Torgo, L. (1997, July) Functional models for regression tree leaves. In: ICML, vol 97, pp 385-393
[3] Chaudhuri, Probal & Huang, M. & Loh, Wei-Yin & Yao, R. (1994) Piecewise-Polynomial Regression Trees. Statistica Sinica 4:143-167
[4] Vogel, D. S., Asparouhov, O., & Scheffer, T. (2007) Scalable look-ahead linear regression trees. In: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, pp 757-764. https://doi.org/10.1145/1281192.1281273
[5] Quinlan, J. R. (1992, November) Learning with continuous classes. In: 5th Australian joint conference on artificial intelligence, vol 92, pp 343-348
[6] Malerba, D., Appice, A., Ceci, M., & Monopoli, M. (2002) Trading-off local versus global effects of regression nodes in model trees. In: Foundations of Intelligent Systems: 13th International Symposium, ISMIS 2002 Lyon, France, June 27–29, 2002 Proceedings 13, pp 393-402. Springer Berlin Heidelberg.
[7] Gadekar, K., & Gore, S. (2019) GENERAL LINEAR REGRESSION. IJRAR-International Journal of Research and Analytical Reviews (IJRAR) 6(1): 42-46
[8] Czajkowski, M., & Kretowski, M. (2016) The role of decision tree representation in regression problems–An evolutionary perspective. Applied soft computing 48:458-475
[9] Dobra, A., & Gehrke, J. (2002, July) SECRET: A scalable linear regression tree algorithm. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pp 481-487
[10] Breiman, L., Friedman, J.H., Olshen, R.A., Stone C.J. (1984) Classification and Regression Trees. Chapman and Hall/CRC
[11] Fisher, R. A. (1988). Iris. UCI Machine Learning Repository. https://doi.org/10.24432/C56C76
[12] Liver Disorders. (1990). UCI Machine Learning Repository. https://doi.org/10.24432/C54G67
[13] Ballabio, Davide, Cassotti, Matteo, Consonni, Viviana, and Todeschini, Roberto. (2019). QSAR fish toxicity. UCI Machine Learning Repository. https://doi.org/10.24432/C5JG7B
[14] Yeh, I-Cheng. (2009). Concrete Slump Test. UCI Machine Learning Repository. https://doi.org/10.24432/C5FG7D
[15] Tfekci, Pnar and Kaya, Heysem. (2014). Combined Cycle Power Plant. UCI Machine Learning Repository. https://doi.org/10.24432/C5002N