Assessment of Machine Learning Techniques Through Accuracy Estimation

[1]G.DIVYA
Research Scholar, Department Of Computer Science,
A.V.V.M Sri Pushpam College (Autonomous), Poondi,Thanjavur(Dt), Affliated to Bharathidasan University,Thiruchirappalli, Tamilnadu (Mail Id-gdivya19.mca@gmail.com)

[2]Dr.V.MANIRAJ
Associate Professor, Research Supervisor, Head of the Department,
Department of Computer Science, A.V.V.M Sri Pushpam College (Autonomous), Poondi,Thanjavur(Dt), Affliated To Bharathidasan University,Thiruchirappalli, Tamilnadu,
(Mail Id-manirajv61@gmail.com)

ABSTRACT

In this paper, we have worked on comparing various data mining algorithms using R tool and various comparison models. After comparison has been done, we have applied the best algorithm as per the result to make the prediction. In this paper, we worked on how to check algorithms on a dataset using R and find the most accurate algorithm model for our dataset. It cannot be easy to tell what algorithm to use on the dataset to get the best results. We don't know the best parameters to use for particular algorithms. Here we worked on the strategy of trial and error to choose the accurate algorithm. Data set is used to train models and a test option is used to evaluate the model. Test metrics are used for comparison. Worked on the various models for which model to choose, to configure them, and pre-process them using data. Applied various techniques for comparing the accuracy of constructed models. We have worked on some algorithms compared them and after choosing one algorithm we can improve the result of algorithms by tuning various algorithms parameters by combining or changing parameters. Once we find the best algorithm applied on the dataset for prediction and tried to improve it by changing various criteria. Here in this paper, we have worked on the decision tree and tried to find out the best-resulting model for prediction. This learning opportunities can be further used in affordable energy, agriculture, and environmentally sound technologies etc.

Keywords*:* SVM, Support Vector Machine; RF, Random Forest; LR, Linear Regression.

1.Introduction

Data Mining is the process that helps us to fine the precious information from large amount of data. Interesting data patterns can be derived from the large size of dataset and dataset are part of each and every business. All business transactions are maintained through the dataset using some random billing software and these datasets are containing large amount of information but problem is how to derive useful part of information from large size of dataset. All information contained in dataset may or may not be useful but data mining can help to extract information using various mining algorithms. Problem here is industrial dataset specially like an e commerce data are large in size and it may contain many useful information to predict the selling of items but due to large size and various number of fields it is difficult to derive useful information. The goal here is to find the best algorithm using various comparison techniques and to implement it on dataset to derive information. There are number of algorithms available in data mining and each one is good at some point and average or bad at some point. In this paper we have used R Tool package to create different data mining algorithms with their parameters and tries to select best algorithm so we can use that to make prediction. Worked on comparison of different algorithms using evaluation metric Kappa and Accuracy. Also prepared data to fit into data mining algorithms and evaluated predictions in R tool. Here we have trained 5 different algorithms with repeated cross validation with 10 folds and 3 repeats and compared. There are many comparison methods and here we have used Summary table, Box and Whiskers plot, density plot and graph and generated decision tree using R tool library to make prediction. Caret R package is used to compare the result of different models [1]. To discover the best algorithm for database we have used trial and error approach. This is like performing an experiment with dataset, comparison and algorithms and we may try various combinations. Apply algorithms on dataset and compare the result of all algorithms once we know which is best for our problem, we can focus on those algorithms to improve the result by changing algorithm parameters Here we have compared various algorithms and as per our result implemented decision tree algorithm using R tool. Also, its trial-and-error approach so this can be improved by comparing more algorithm and comparison technique [3]. This paper contributes which model to choose and how to configure and pre-processing of data and after that comparing the accuracy and estimate the models using evaluation metrics.

1.1 Related works

[6a,b] Here authors have used Machine Learning Techniques like SVM, CART, DART, Linear Regression, Random Forest and Neural Networks for detecting phishing in emails. They have very well discussed error rate of every technique used for the detection based on Precision, Recall, and f1 for every classifier and it's found that RF has the lowest error rate followed by CART, LR, BART, SVM and NN. Another measure considered is ORC curve that shows plot between false positives and true positives and it gives a better result. This shows that NN has highest accuracy followed by RF, SVM, LR, BART and CART. 10- Fold cross validation had been used and averaged the average. The final discussion from the paper concludes that there cannot be one fixed measure for comparing the classifiers but it depends on the usage and type of data we are applying the techniques for our results. In addition, we need to note that error is considered without applying costs does not provide the right way of working with it. [12b] The paper is about network management and surveillance in regard with traffic control systems. The study had been carried on packet length and interval arrival time distribution. Parameters considered for comparison of machine learning techniques are correlation and consistent based for reducing the set feature. The ML techniques discussed are Bayesian Network, Naïve Bayes, C4.5. Computational performance is shown as one of the most important features to compare the algorithms along with classification accuracy as computational performance effects the accuracy highly. There are many feature reduction algorithms like greedy search, Best Fit Search, Correlation and Consistency. However, here they have discussed the latter two. WEKA Tool had been used for performing the experiments and generate the result. It had been found that feature reduction techniques have the capacity to reduce feature space, which significantly increase computation performance with low impact on classification accuracy. [5] a new version of ID3 have been proposed and the result shows that improved ID3 perform well with less number or leaves and high predictive accuracy for decision tree models. The disadvantages of ID3 have been overcome for attribute selection criteria. The proposed method increases the ranking information between mutual candidate and decision candidate and decreases the same between candidate and selected conditional attributes of the same branch. Improved ID3 calculates the close contact between attributes and decision attributes. It makes sure that for every iteration it selects important attribute and not more attributes like actual ID3 algorithm. Hence, it improves the accuracy and decreases number of leaf. [7] A new decision tree called CCPDT have been proposed that means Class Confidence Proportion Decision Tree. It is robust, effected by size of class, and generated statistically significant rules. To implement CCP top-down and bottom-up approach had been used to prune branches of the tree, which are not relevant. The result is confirmed using CART. C4.5, SPARCC

and HDDT. Reason of creating imbalance in decision tree because of number of classes is discussed as information Gain is sensitive to class imbalance and then it's been solved using CCP-Class confidence proportion. Information gain is shown with confidence of rule. And the results carried out shows that accuracy is improved along with Fisher exact Test is based on statistically significant rules. [8] Worked on K-Means Clustering and C4.5 Decision tree algorithm and proposed a supervised network anomaly detection method. Proposed algorithm is used to detect anomalies presented in the supervised dataset. Classification is used to learn the model and classify a test instance. Classification based techniques is divided into two phases, training and testing phase. Author worked on evaluate the performance of K-Means and C4.5 classifier and compare it using different performance measures. The K-means is used to partition the training instance and C4.5 is used to build the decision tree. Decision tree refines the decision boundary and for detail final conclusion final result is exploited. [9] Here authors have worked on several Supervised Machine earning algorithms and determines most efficient classification algorithm based on the dataset. Different Machine Learning algorithms were considered using WEKA tool. The comparative analysis is done between various supervised algorithms using different attributes of dataset. In order to predict the accuracy and ensure precision for different machine learning algorithms, this research work was carried out by tuning the parameters with two different sets of number of instances. Mean Absolute Error is used to measure forecast and prediction and Kappa Statistic is used to observe accuracy. Also, they found that best algorithm for a particular dataset does not guarantee the accuracy and prediction but under which condition a particular method can perform better. It may vary on attributes or conditions and correlation among the variables make more efficient output models. [10] Authors worked on comparison of Machine Learning algorithm and performed comparative analysis by checking efficiency of algorithms. Here two different datasets are used with 10-cross validation strategy. Outcomes explains that it requires preparation of the model for classifier. [2] Here authors worked on machine learning technique and applied for the fruit image classification and prediction from large dataset where five models are developed and performance are compared to predict the fruit name. They have applied supervised machine algorithm for fruit image classification and prediction. Proposed work mainly consists of five steps like loading and processing data, analysis of data, training the models and result evaluation. (Cernak, 2010) Authors have used Decision Tree as one of the most effective tools for improving and evaluating speech recognition process complexity based on findings and usage of DT for various research work. Majorly 5 different variations are used namely as CART, C4.5, MML, Strict MML and Bayesian Tree. And after experimenting and evaluating the result it's concluded that CART have performed better than other 4 methods selected. The OLLO database is taken for implementing DT based on the feature selection of speaking rate, speaking effort, style. And it's found that CART outperforms C4. Entropy and splitting have been considered as best properties for selecting CART as success method in the experiment. (Wiyono & Abidin, January 2019) A study had been done with students' data to find non active students so that the overall result is not effected. Prediction models as Decision Tree, SVM and KNN is taken based on literature survey and data considered. Authors have found that after experimenting and predicting the results Support Vector Machine performs best with 95% accuracy followed by Decision Tree with 93% and lastly KNN with 92% accuracy. Here combination of Continuous and Categorical data is used. The main attributes of data used are Grade Point, GPA, Hometown, Type of school, major and parent's job, ATKTIF. Training and test data is considered in 75-25% ration with 10-fold cross validation repeated 3 times. Finally result is compared based on confusion matrix generated all three models considered. (AbdulHussien, 2017) Web mining allows us to extract useful information from web using some tools or methods. Classifying these data is a concern for research. In this paper authors have worked upon supervised learning algorithms to classify web data into predefined categories of web documents. Artificial Neural Network, Random Forest and AdaBoost are used behavior comparison for problems faced in classification of web pages. Entropy is used for analysis of text. Experiment is based upon showing effectiveness and difference of algorithm. Data set is divided into 70-30% training and test data. Pages considered as data are divided into 8 categories: Heart. Liver, Respiratory etc. Result is evaluated on the basis of classification Performance using F- Measure considering Precision and Recall. The experiment result shows that accuracy of Random Forest is better that other two methods with 87% accuracy followed by Artificial Neural Network- 84% accuracy and minimum accuracy of Adobos with 81.7%. (Anyanwu & Shiva) Data Mining and Classification are co related to each other these days I research of various topics like fraud detection, AI, credit card rating, customer retention etc. and Decision Tree is one of the most commonly used method as it is simple and effective. In this paper authors have discussed about various serial Decision Tree algorithms and evaluated their efficiency and proficiency with testing of some data set. Here SPRINT, SLIQ, IDE3, C.4.5 and CART had been evaluated in detail to understand the classifiers on health, science, transportation and financial data. Parameters considered to evaluate performance of classifiers are number of records and attributes along with size of class. Accuracy of model is studied based on execution time of the data to generate prediction. Experiment explains that there is a relation between size of data set and time of execution. Also, SPRINT and CART has a good accuracy in classification as compared other methods used. Future scope can be performing the experiment on parallel implementation of tree algorithms and compare the result with serial methods. [11,12]This paper discusses about when modelling or algorithmic style is used for supervised and unsupervised learning. Both the methods are discussed for distinguished task at edge and clouds for the networking purpose at different layers of protocols in the physical layer specifically. Concept and Implementation of supervised and un-supervised learning process has been discussed here with its detailed understanding and application. How ML can be used an error tolerance method and how data can be an important phase but should not be the only way to implement learning models but used to aid to the results and understand the pattern. [4] New way of comparing learning algorithms had been introduced that works with non-parametric test such as poison binomial test. Dependencies induced for classification is evaluated using Bayesian Algorithm. The result is compared and better than Wilcoxon signed rank test and signed test. ROC curve is used to obtain desired AUC a trapezoidal approximation of the values for success and error count when threshold is minimum. [12a] This paper is for medical work and specifically specifies that machine learning had started becoming an important part for disease prediction. And more over authors here tried to analyses the key trends in supervised learning algorithms. Two medical databases Scopus and PubMed were considered for item search. And the result shows that Support Vector Machine is mostly applied algorithms followed by Naive Bayes. Random Forest algorithm has highest accuracy test followed by SVM. The study has been done for many algorithms like SVM, KNN, DT, Naive Bayes, Logistic Regression, ANN, and Random Forest. Confusion matrix and ROC Curves were considered for result comparison. (Wadhwa, Shivani and Saluja, Kamal, 2022) proposed a hybrid approach using Capsule network and MultiLayer Perceptron (MLP) has been used for classification of images. Ethereum blockchain platform is used for providing security to the global data. The proposed framework provides more accuracy and security for the detection of COVID-19 patients and monitoring their health [13–19]. (Kusum Yadav, Sunil Gupta, 2020) uses Euclidian distance formula as well Mehlanobis to find the minimum distance between slots as technically we called as clusters, and we have also applied the same approach to ant colony optimization (ACO) algorithm which results in the production of two- and multi-dimensional matrix [20].

2. Methodology

Here we are going to use Watson Analytics Sales Database which is sales product database and will use it to make comparison of algorithm. For implementation we have used R tool that has various library functions that we can use to implement the algorithms. Dataset used for experimenting the result has various columns like Order method type, Retailer type, Product line, Product type, Product, Year, Quarter, Revenue, Quantity, Gross margin. This dataset contains more than 88 thousand rows of data and 111 column and for implementing in R tool we have broken down the dataset in to smear size to understand the result. Also Pre-processing of the data is done as it is one of the important things to perform on the data before implementing any algorithm. By pre-processing of the data, we have tried to settle unwanted data like empty fields, missing data, wrong formatted data and noise. Various Data mining algorithms are available and when we need to choose the best algorithm which is most accurate it is difficult due to different characteristics of algorithm. The goal here is to choose most accurate model but it is difficult to select one algorithm to solve the problem. So to overcome this we have used trial and compare approach and we have studied the result of various algorithm using our same database which we have described earlier. We have finalized the database with limited amount of data and implemented various data mining algorithms using R tool and now we are going to compare the result of all. We have selected some algorithms that may fit our work as they have different characteristics.

They are:

- Classification and Regression Tree
- Support Vector Machine
- K-Nearest Neighbour algorithm
- Random Forest
- Naïve Bayes

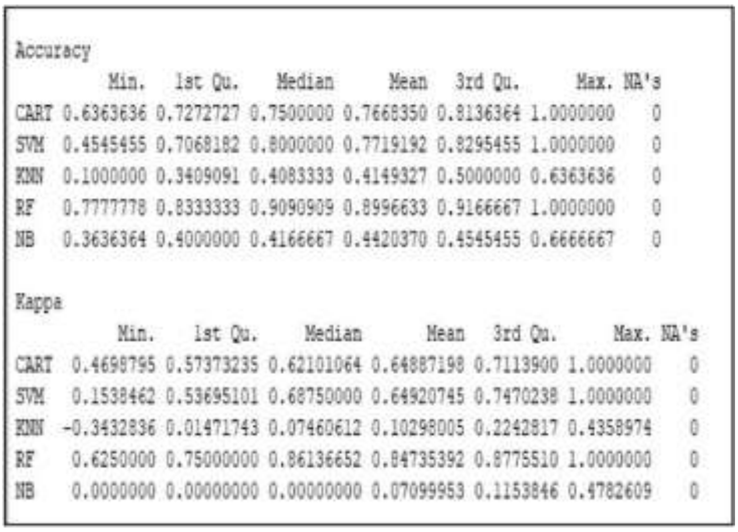
Repeated cross validation is used as common configuration standard, we use 10 folds with repeats. Evaluation matric used are accuracy and Kappa. Pre-processing is required to make data suitable for implementing data to data mining techniques. All attributes we are not going to use so we have to use is wisely which may affect the result. There are many approaches of data pre-processing with different purpose to process as real data may noisy and incomplete. We are going to split the data in to two parts Training Set and Test Set. Training set us used to represent sample that we use for model creation and test set is used for performance. We split the dataset in ration of 70:30 where 70% used in train and 30% used in test. Using Training dataset, we fit the model and using Test dataset we provide the final model. Splitting the dataset depends on 2 things that is size of the data and actual model we are training. We run the model against the test data, calculate the result and compare it with expected result to get the accuracy of the model. We have used Accuracy and Kappa evaluation metrics to compare the accuracy. Here Accuracy is percentage of correct classified upon total classified instance. Kappa is normalized at baseline of random selection of the dataset.

3. Result and discussion

Table Summary is easiest comparison done by using summary function. It creates a table with one algorithm for each row and evaluation metrics for each column. In Fig. 1 we can see the Summary Data. This includes min, median and max values and also some percentile for each metrics. The Kappa values can be evaluated as Poor agreement that is less than 0.20 and it goes up to Very good agreement that is 0.80–1.00. Based on this we can say RF is very good agreement as it has good result for our database. CART and SVM seems just after RF and KNN has the poor agreement for Kappa and moderated for Accuracy. Table 1 is the summary table derived from the original result (see Table 2).

3.1. The Density Plots

Density plots shows the distribution of the data over time periods with similar behavior like histogram. We can compare the distribution of variables for various data mining algorithms.



```
Accuracy
          Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
CART 0.6363636 0.7272727 0.7500000 0.7668350 0.8136364 1.0000000    0
SVM  0.4545455 0.7068182 0.8000000 0.7719192 0.8295455 1.0000000    0
KNN  0.1000000 0.3409091 0.4083333 0.4149327 0.5000000 0.6363636    0
RF   0.7777778 0.8333333 0.9090909 0.8996633 0.9166667 1.0000000    0
NB   0.3636364 0.4000000 0.4166667 0.4420370 0.4545455 0.6666667    0

Kappa
           Min.    1st Qu.     Median       Mean   3rd Qu.      Max. NA's
CART  0.4698795 0.57373235 0.62101064 0.64887198 0.7113900 1.0000000    0
SVM   0.1538462 0.53695101 0.68750000 0.64920745 0.7470238 1.0000000    0
KNN  -0.3432836 0.01471743 0.07460612 0.10298005 0.2242817 0.4358974    0
RF    0.6250000 0.75000000 0.86136652 0.84735392 0.8775510 1.0000000    0
NB    0.0000000 0.00000000 0.00000000 0.07099953 0.1153846 0.4782609    0
```

Fig.1. Summary data

Table1
Summary table.

Technique	Kappa	Inference	Accuracy	Inference
CART	0.573–0.7113	Good Agreement	0.727–0.813	Good Agreement
SVM	0.536–0.747	Good Agreement	0.706–0.829	VeryGood
KNN	—0.014–0.224	Fair Agreement	0.340–0.500	Moderate Agreement
RF	0.750–0.877	VeryGood	0.833–0.916	VeryGood
NB	0.000–0.115	Poor Agreement	0.400–0.4545	Moderate Agreement

Table 2 Summary target variable and criterion.

Target Variable	Accuracy	Complexity Parameter	Criterion
V3	36.14	0.01297648	Information
V3	31.33	0.01297648	GiniIndex
V4	54.76	0	Information
V4	55.95	0.006111111	GiniIndex

With multiple colors we can represent multiple algorithms at same time with continuous line presentation for points plotted for various values using dataset. We can see in Fig. 3 that CART, KNN, NB, RF and SVM are presented with different colors and different coverage area that shows distribution. The points at the bottom shows the intensity of data and data at peaks represents the mean value for each model. Here KNN is correlated to NB and SVM has less accuracy and lesser Kappa compared to CART and RF. CART and RF present good distribution of data with some common values. RF has three peaks that shows good distribution of different section.

3.2. The box and whisker plots

It is lots of each individual variables where boxes are ordered from highest to lowest mean accuracy. It is useful to look at the mean values that is dots and it overlaps the boxes. It is effective to check the distribution of the data. The extreme values at any end of the scale are sometimes included on the display to show how they are extended beyond majority of the data. Boxes in figure are ordered from highest to lowest mean accuracy. The left is 25% and right is 75% with median is always 50%. Anything that is present outside the whiskers are outliers and here we can see NB is not a good choice to generate the tree. Skewness is degree of distortion of data from central value r. It can be understood by shape of the box plots. Here RF and CART have normal distribution and NB has not good form of skewness. From Fig. 2 this result we can predict that RF and CART are looking more effective. *3.3. The Statistical Significance Tests* This is helpful to find the relations between algorithms and to verify result so it helps to validate our data and result before we conclude any result. It is going to help us to decide that which classification technique we are going to use for forming a decision tree. It has two key variables effect and sample size where Effect is the difference between the results of two sample sets and sample size is about size of data considered for test. We can calculate the significance of the difference the matrix distributions of various algorithms. Here lower diagonal of the table shows p-values and upper diagonal of table shows different between distributions. It helps to check and compare accuracy between different models. P-value is used to indicate the probability of uncorrelated systems. It is probability of relations showing that two variables has no relation. Here in Fig. 4 p-value is near to 0 or less than 0.05 that shows significance relation between CART, KNN, RF and NB. Here we reject null hypothesis, H0 and accept the H1. Here we have considered null hypothesis H0 that says that. **H0**. there is no significance relationship between CART, SVM, KNN, RF and NB. H1: There is a relationship between CART, SVM, KNN, RF and NB. So, from this we can use CART or RF for forming decision tree for our work. Also, we are more focused on CART as it generates detailed tree whereas RF generates summarized decision tree.

4. Results-plotting of graphs As we have used limited number of rows and pre-processing already has been applied using R. We have used train () method for training the dataset for different algorithms. From above comparisons we have found the CART is possible giving best result so we have implemented decision tree. We can implement this with different set of variables that is columns here with respect to dataset we are using.

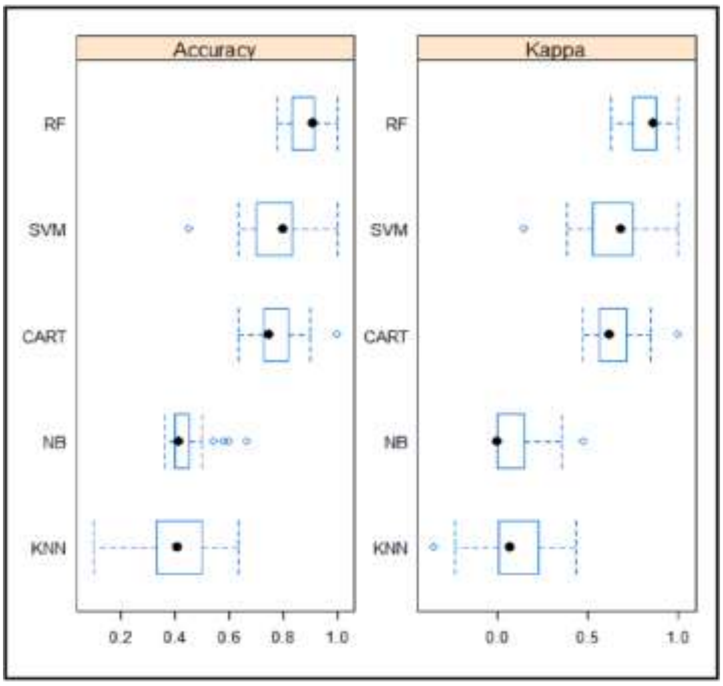


Fig. 2. Whiskers plot.

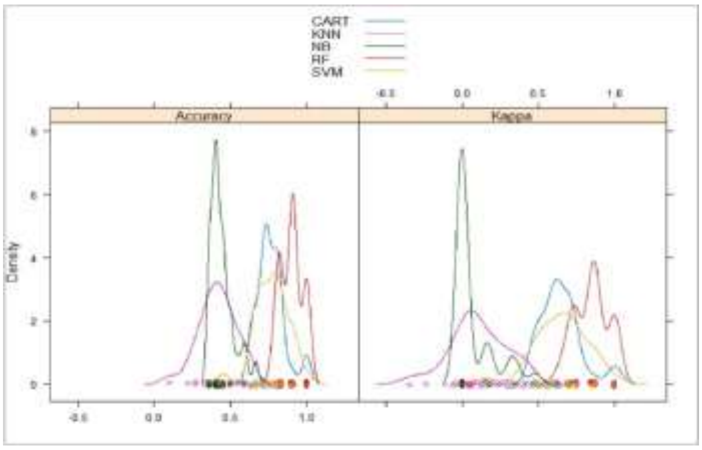
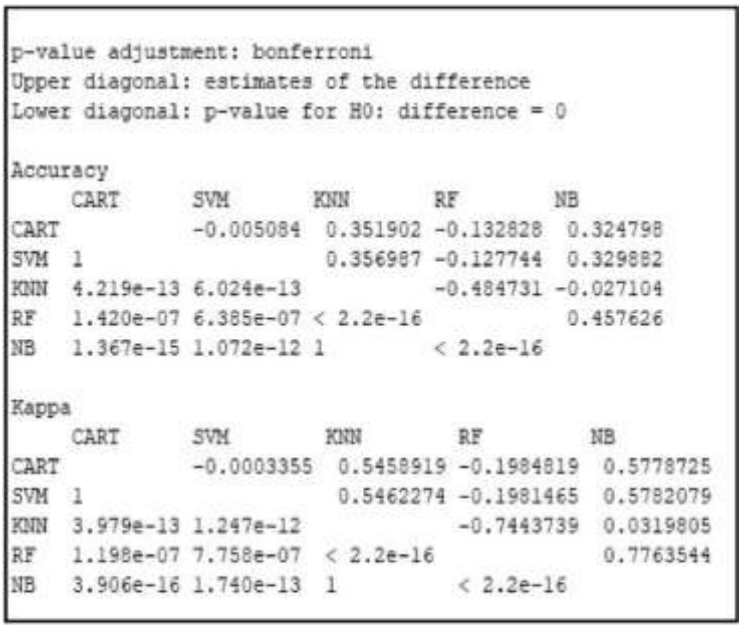


Fig. 3. Density plots.



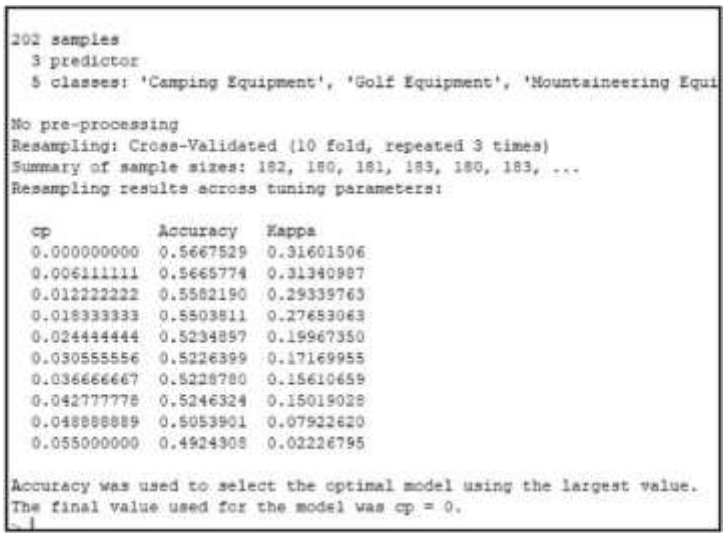
```
p-value adjustment: bonferroni
Upper diagonal: estimates of the difference
Lower diagonal: p-value for H0: difference = 0

Accuracy
     CART      SVM       KNN       RF        NB
CART           -0.005084  0.351902 -0.132828  0.324798
SVM  1                    0.356987 -0.127744  0.329882
KNN  4.219e-13 6.024e-13           -0.484731 -0.027104
RF   1.420e-07 6.385e-07 < 2.2e-16            0.457626
NB   1.367e-15 1.072e-12 1         < 2.2e-16

Kappa
     CART      SVM        KNN        RF         NB
CART           -0.0003355  0.5458919 -0.1984819  0.5778725
SVM  1                     0.5462274 -0.1981465  0.5782079
KNN  3.979e-13 1.247e-12             -0.7443739  0.0319805
RF   1.198e-07 7.758e-07  < 2.2e-16              0.7763544
NB   3.906e-16 1.740e-13  1          < 2.2e-16
```

Fig. 4. Statistical significance tests.

We can implement this with different set of variables that is columns here with respect to dataset we are using. We have selected one target variable and implemented decision tree using rpart package of R tool. For variation and to follow the approach of trial and compare we have used different target variable but here we are going to discuss the best result possible. So here we have used variables V4, V1, V2, V9 and V10 where V4 is set to our target variable and details are V4 is for “product line”, V1 presents country “retailer country”, V2 is for “order method type”, V9 is for “revenue” and V10 is for “quantity” This is the database we have already describe earlier. Also, for the splitting criteria we have used spit parameter with two different values, one value is “information” which is used for information gain and other value is “gini” which is used for gini index. Here in Fig. 5, we can see CP value as 0.006 so tree may be derived as complex. Fig. 6 shows Decision tree with CP value as 0.006, Target variable V4 and Split parameter as Gini. It shows accuracy is 55.95 for the test set and we have described the Gini here as with compared to other criteria it is giving the best result.



```
202 samples
  3 predictor
  5 classes: 'Camping Equipment', 'Golf Equipment', 'Mountaineering Equi

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 182, 180, 181, 183, 180, 183, ...
Resampling results across tuning parameters:

  cp           Accuracy   Kappa
  0.000000000  0.5667529  0.31601506
  0.006111111  0.5665774  0.31340987
  0.012222222  0.5582190  0.29339763
  0.018333333  0.5503811  0.27653063
  0.024444444  0.5234897  0.19967350
  0.030555556  0.5226399  0.17169955
  0.036666667  0.5228780  0.15610659
  0.042777778  0.5246324  0.15019028
  0.048888889  0.5053901  0.07922620
  0.055000000  0.4924308  0.02226795

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was cp = 0.
>
```

Fig. 5. V4 (product line) as Target Variable and Split Parameter as Gini.

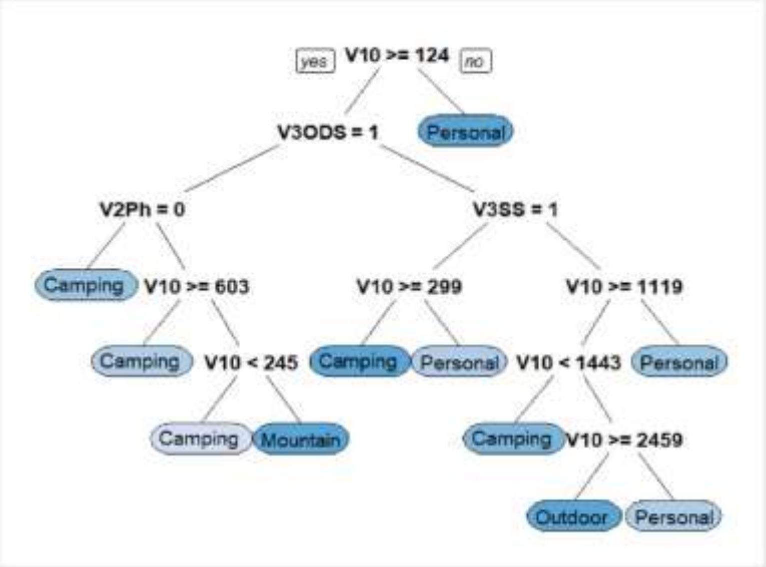


Fig. 6. Decision tree with V4 as Target Variable and Split Parameter as Gini.

In similar way we can change the target variables and split criterion as Information and Gini index and on the trail and Compare bases result can be derived.

5. Conclusion

After comparison of various algorithm, we have found that CART is giving best result as per our result and data. But result me varies on database, size and other aspects as well. Here we can see that certain combination of variable and criterion can give the accurate result. Here target variable V4 with criterion gini index gives the highest accuracy rate. We started by comparing various algorithms using estimated accuracy of the constructed models and for that we prepared data trained models and used evaluation metrics. Using this result a decision tree can be derived and we make the prediction using the result. We worked on trial-and-error approach and continue the process till we get the best result. We can derive different result on different combinations so this process can be repeated till we found the best result. Also, dataset, size, fields and many other aspect can impact the result. Steps begin with which model to choose and how to configure and pre-processing of data and after that comparing the accuracy and estimate the models using evaluation metrics. For future work we can work with a greater number of data using comparison methods we have used to get more accuracy in the result so final output could became more informative.

References

[1] Ansam A. Abdul Hussien, Comparison of machine learning algorithms to classify web pages. (ijacsa), Int. J. Adv. Comput. Sci. Appl. 8 (11) (2017), https://doi.org/ 10.14569/IJACSA.2017.081127.

[2] R. Agarwal, P. Sagar, A comparative study of supervised machine learning algorithms for fruit prediction, J. Web Dev. Web Des. 4 (1) (2019) 14–18.

[3] Anyanwu, M., & Shiva , S. (n.d.). Comparative analysis of serial decision tree classification algorithms. Int. J. Comput. Sci. Secur., Volume (3), Issue (3).

[4] A.L. Fran, c. Marchand, Bayesian comparison of machine learning algorithms on single and multiple datasets, in: Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS), 22, W&CP 22, La Palma, Canary Islands, 2012 n.d.

[5] Liang, X., Qu, F., & Yang , Y. (n.d.). An improved ID3 decision tree algorithm based on attribute weighted. International Conference on Civil, Materials and Environmental Sciences (CMES 2015). Published by Atlantis Press.

[6] a S. Abu-Nimeh, D. Nappa, X. Wang, S. Nair, A comparison of machine learning techniques for phishing detection, APWG eCrime Researchers Summit (2007); b P.A. Pittsburg, M. Cernak, A comparison of decision tree classifiers for automatic diagnosis of speech recognition errors, Comput. Inf. 29 (2010) 489–501.

[7] Liu, W., Chawla , S., Cieslak, D. A., & Chawla, N. V. (n.d). A Robust Decision Tree Algorithm for Imbalanced Data Sets. SIAM.

[8] A.P. Muniyandia, R. Rajeswarib, R. Rajaramc, Network anomaly detection by cascading K-means clustering and C4.5 decision tree algorithm, Int. Conf. Commun. Technol. Syst. Des. (2011) 174–182.

[9] F.Y. Osisanwo, J. Akinsola, O. Awodele, J.O. Hinmikaiye, O. Olakanmi, J. Akinjobi, June, Supervised machine learning algorithms: classification and comparison, Int. J. Comput. Trends Technol. (IJCTT) 48 (3) (2017) 128–138.

[10] K. Sethi, A. Gupta, G. Gupta, V. Jaiswal, Comparative analysis of machine learning algorithms on different datasets, in: International Conference on Innovations in Computing (ICIC 2017), Simeone, O., 2017, pp. 87–91, https://doi.org/10.1109/ TCCN.2018.2881442. Dec. 2018.

[11] A very brief introduction to machine learning with applications to communication systems. IEEE Trans. Cognit. Commun. Netw., vol. 4 (no. 4), 648-664.

[12] a S. Uddin, A. Khan, M. Hossain, M. Moni, Comparing different supervised machine learning algorithms for disease prediction, BMC Med. Inf. Decis. Making (2019); b N. Williams, S. Zander, G. Armitage, A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification, Comput. Commun. Rev. 36 (5) (2006).

[13] s. Wiyono, t. Abidin, Comparative study of machine learning knn, svm, and decision tree algorithm to predict student's PERFORMANCE, Int. J. Regul. Govern. 7 (1) (January 2019) 2394–3629.

[14] Kusum Yadav, Sunil Gupta, Hybridization of K-Means Clustering Using Different Distance Function to Find the Distance Among Dataset, Springer Science and Business Media LLC, 2021.

[15] S. Chordia Anita, S. Gupta, An effective model for anomaly IDS to improve the efficiency, in: 2015 International Conference on Green Computing and Internet of Things, ICGCIoT), 2015, pp. 190–194, https://doi.org/10.1109/ ICGCIoT.2015.7380455.

[16] K. Saluja, A. Bansal, A. Vajpaye, S. Gupta, A. Anand, Efficient bag of deep visual words based features to classify CRC images for colorectal tumor diagnosis, in: 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering, ICACITE), 2022, pp. 1814–1818, https://doi.org/ 10.1109/ICACITE53722.2022.9823727.

[17] B.K. Sahoo, A. Sardana, V. Solanki, S. Gupta, K. Saluja, Novel approach of diagnosing significant metrics of load balancing using CloudSim, in: 2022 10th International Conference on Emerging Trends in Engineering and Technology - Signal and Information Processing, ICETET-SIP-22, 2022, pp. 1–6, 10.1109/ ICETET-SIP-2254415.2022.9791688.

[18] S. Gupta, M. Shahid, A. Goyal, R.K. Saxena, K. Saluja, Black hole detection and prevention using digital signature and SEP in MANET, in: 2022 10th International Conference on Emerging Trends in Engineering and Technology - Signal and Information Processing, 2022, pp. 1–5, https://doi.org/10.1109/ICETET-SIP-2254415.2022.9791738 (ICETET-SIP-22).

[19] Wadhwa, Shivani and Saluja, Kamal and Gupta, Sunil and Gupta, Divya, Blockchain based Federated Learning approach for Detection of COVID- 19 using Io MT (July 14, 2022). Available at SSRN: https://ssrn.com/abstract=4159195 or https://doi.org/10.2139/ssrn.4159195.

[20] K. Yadav, S. Gupta, N. Gupta, S.L. Gupta, G. Khandelwal, Hybridization of K-means clustering using different distance function to find the distance among dataset, in: T. Senjyu, P.N. Mahalle, T. Perumal, A. Joshi (Eds.), Information and Communication Technology for Intelligent Systems. ICTIS 2020. Smart Innovation, Systems and Technologies, 195, Springer, Singapore, 2021, https://doi.org/ 10.1007/978-981-15-7078-0_29.