A Comparative Study on Car Evaluation dataset using different machine-learning methods

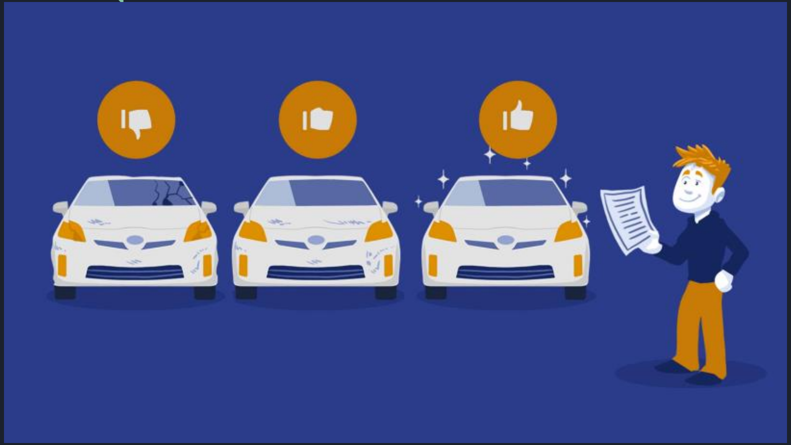
Group member: Shiyu Wang, Yingda Tao, Hao Zeng, Yuqing Jin

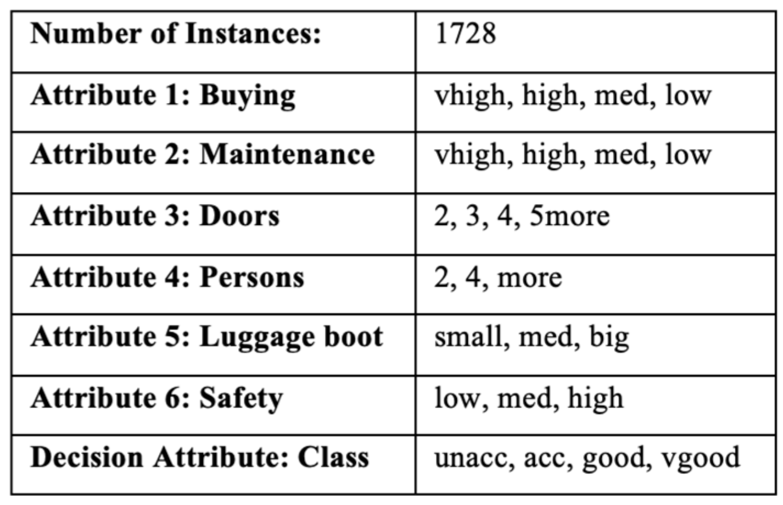
Body of Representation

- Dataset Description
- Reproduction (NB & MLP-ANN)
- Other ML methods (SVM, KNN, Decision Tree, Random Forest, etc.)
- Conclusion

Dataset Description

Data Description



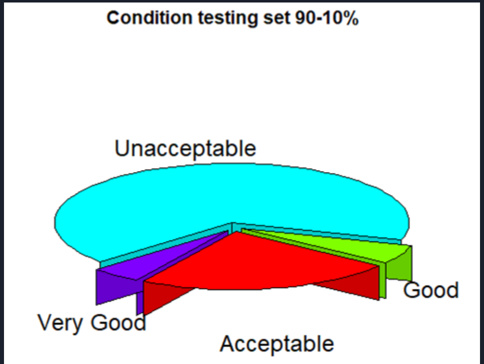


Number of Instances:	1728
Attribute 1: Buying	vhigh, high, med, low
Attribute 2: Maintenance	vhigh, high, med, low
Attribute 3: Doors	2, 3, 4, 5more
Attribute 4: Persons	2, 4, more
Attribute 5: Luggage boot	small, med, big
Attribute 6: Safety	low, med, high
Decision Attribute: Class	unacc, acc, good, vgood

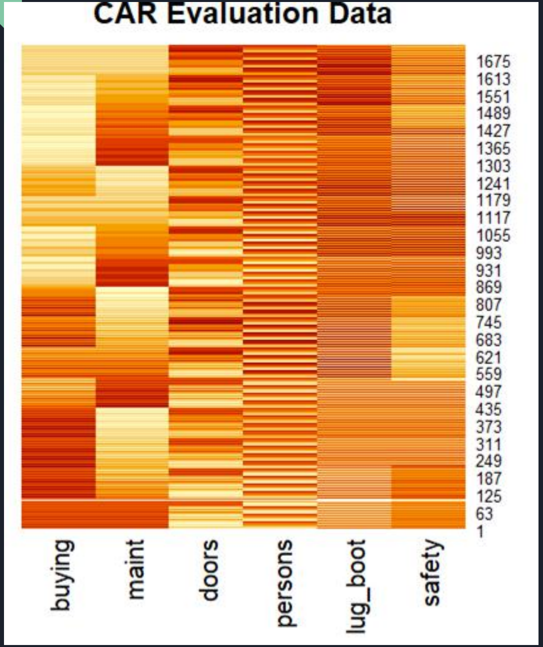
People would make an evaluation before buying a car. They would consider the price, maintenance, seat number, door number and luggage boot etc.

Six predictors and one response in car evaluation dataset. Total 1728 instances in the dataset.

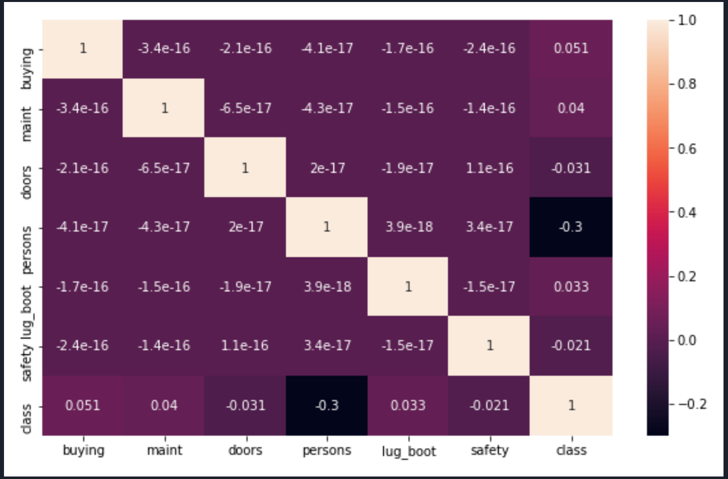
Data Visualization



Pie chart of the decision attribute (class).
Most of the instances belong to ‘unacceptable’.
‘Good’ and ‘very good’ are far less than others.



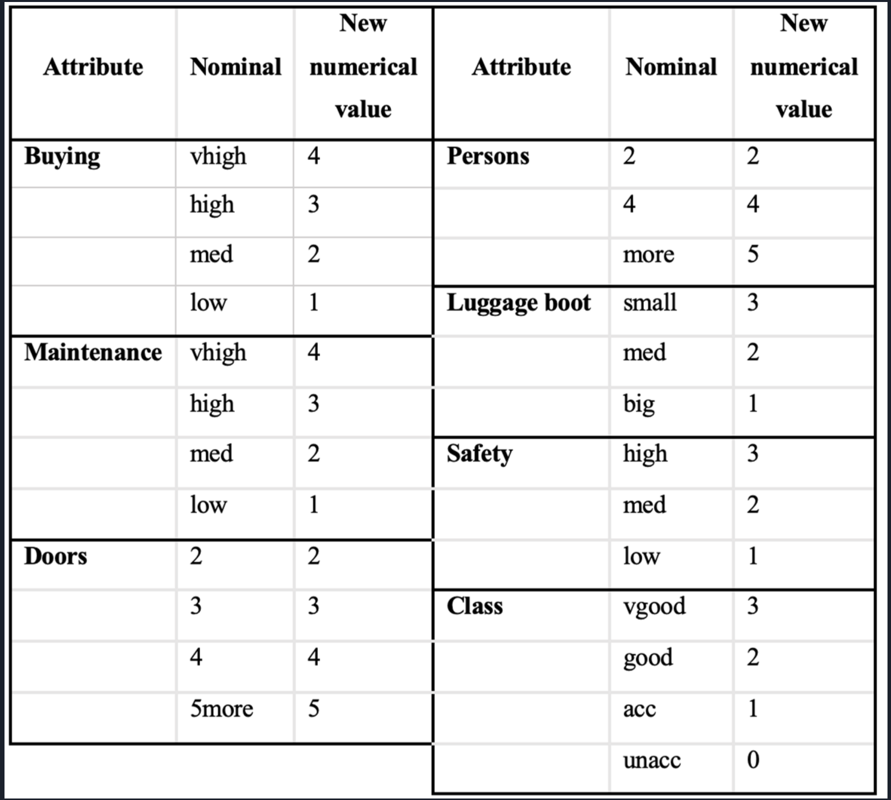
Pheatmap (clustered heatmap) of our six attributes. From the graph, we could see how dispersed each of attribute is.



Correlation matrix of all the attributes. Correlation coefficients are denoted on the graph.

Data Preprocessing

- Data-Cleaning:
  Convert nominal attributes to numerical value
- Min-Max Normalization:
  Narrow data between 0-1
- Data-Split:
  Training : Testing =
  90%-10%, 66%-34%, 50%-50%, and 10-fold cross validation



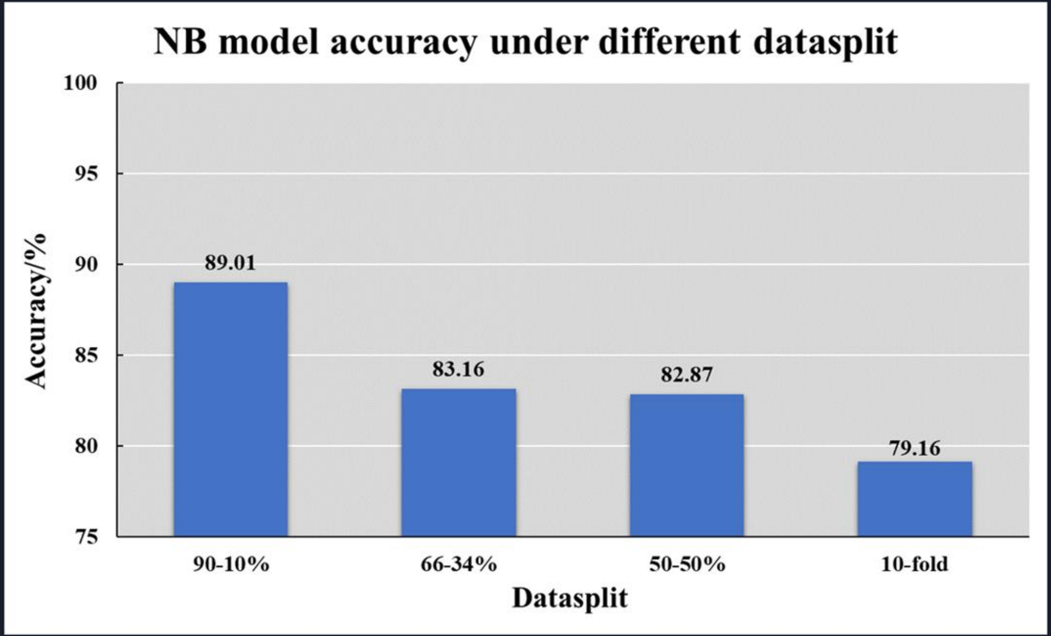
Attribute	Nominal	New numerical value
Buying	vhigh	4
	high	3
	med	2
	low	1
Maintenance	vhigh	4
	high	3
	med	2
	low	1
Doors	2	2
	3	3
	4	4
	5more	5
Persons	2	2
	4	4
	more	5
Luggage boot	small	3
	med	2
	big	1
Safety	high	3
	med	2
	low	1
Class	vgood	3
	good	2
	acc	1
	unacc	0

Change 6 predictors and 1 decision attribute to numerical value, to facilitate data analysis.

Reproduction

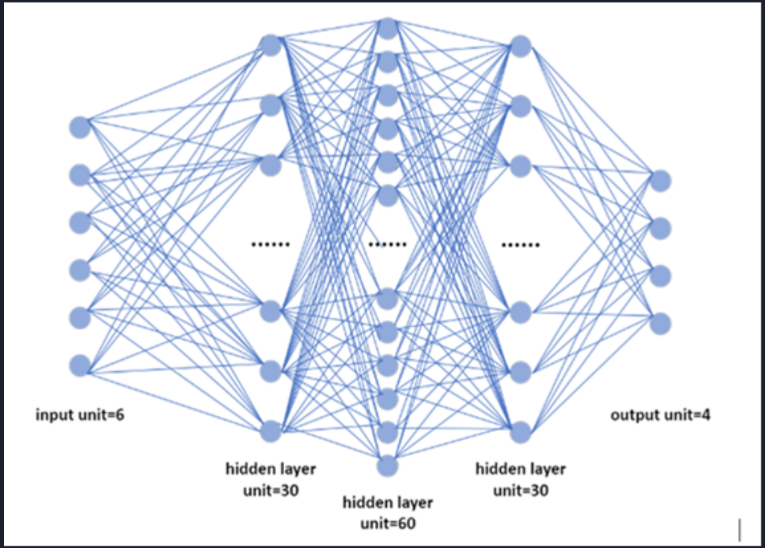
Naive Bayes

$$\hat{f}(x) = \underset{y \in Y}{\operatorname{argmax}} P[X = x | Y = y] * P[Y = y] = \underset{y \in Y}{\operatorname{argmax}} \prod_{j=1}^{d} P[X_j = x_j | Y = y] * P[Y = y]$$



The structure of ANN

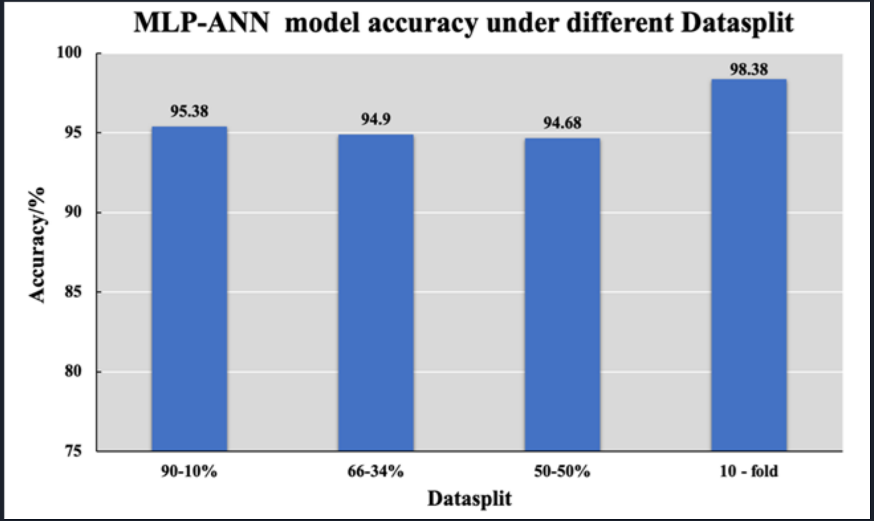
MLP-ANN



The structure of ANN

NB model accuracy in the original paper

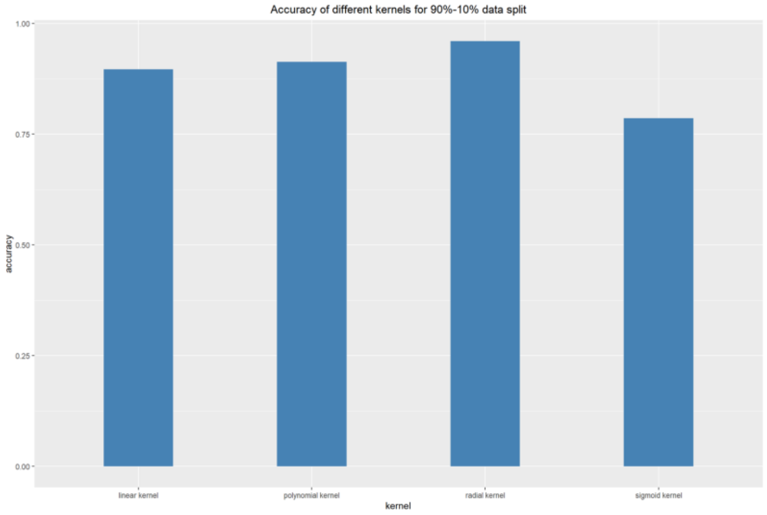
Data Split	Accuracy
90-10%	94.79%
66-34%	93.19%
50-50%	92.70%
10-fold	94.09%



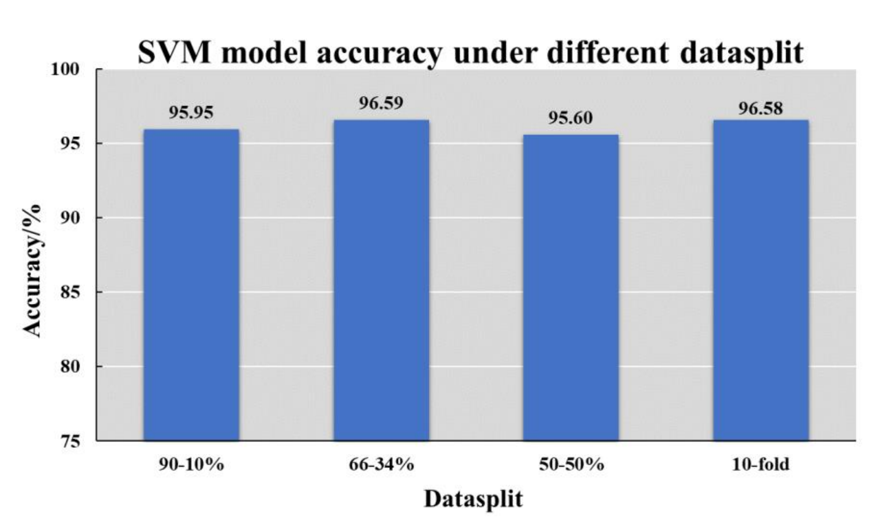
Other ML methods

SVM[3]

SVM using kernel trick

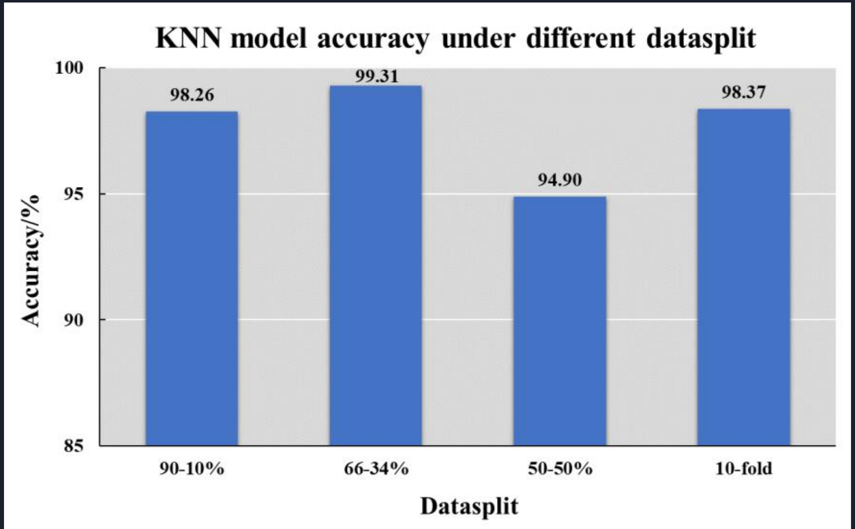


Accuracy of SVM model using radial kernel



KNN[4]

Accuracy of KNN model with K=5

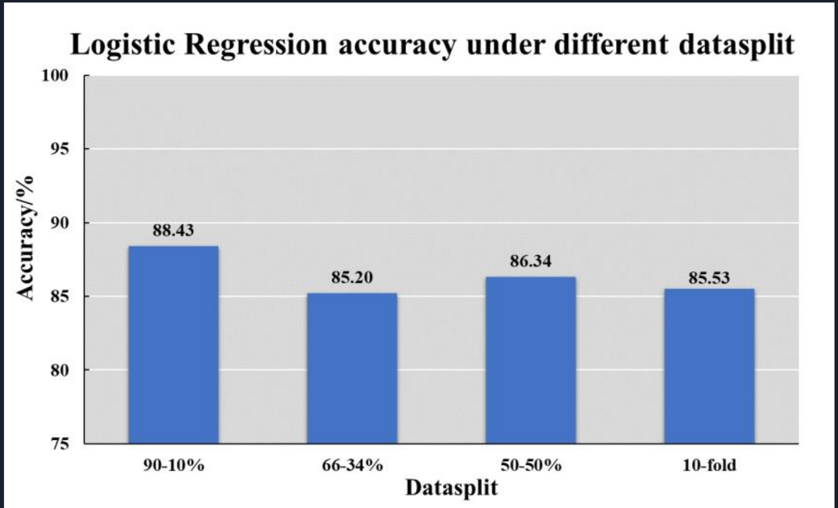


When k is too small -> sensitive to noise (overfitting problem)
When k is too large -> bad accuracy (k=200, accuracy decreases to 80%!)

Logistic Regression

Not regression method, but a classification method [5]

Accuracy of Logistic Regression

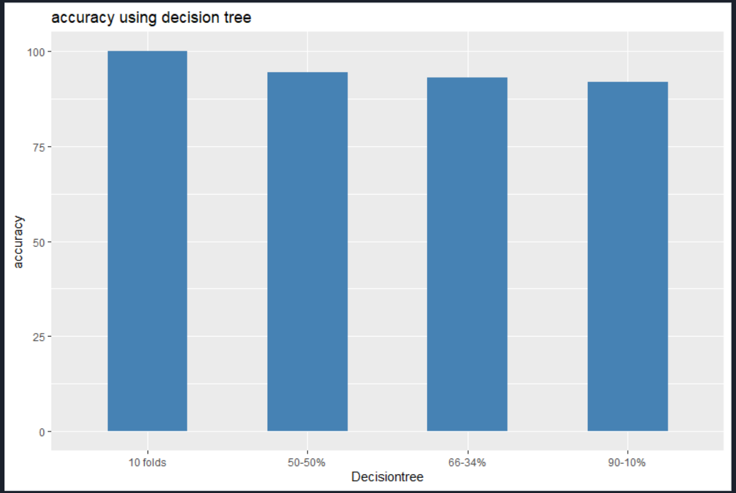


Result is not good: Logistic Regression method is usually used for solving binary classification problems, not suitable for multi-classification problem.

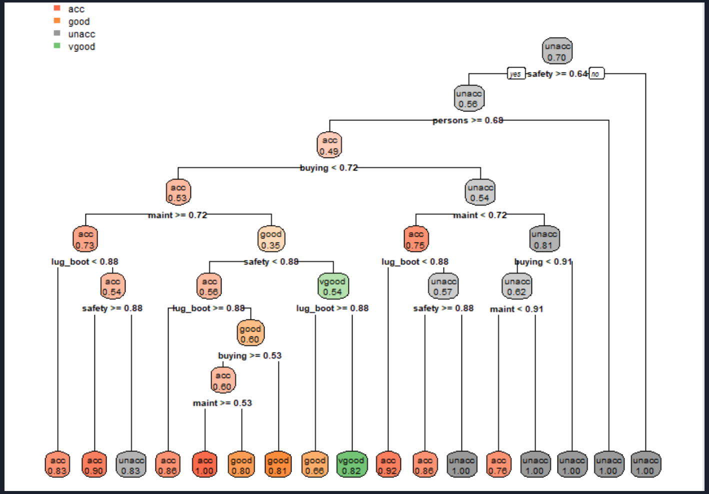
Decision Tree Classification[1]

- Decision tree is a nonparametric method.
- It can be used to compare with parametric methods (such as naive Bayes)
- CART algorithm

Accuracy of decision tree method



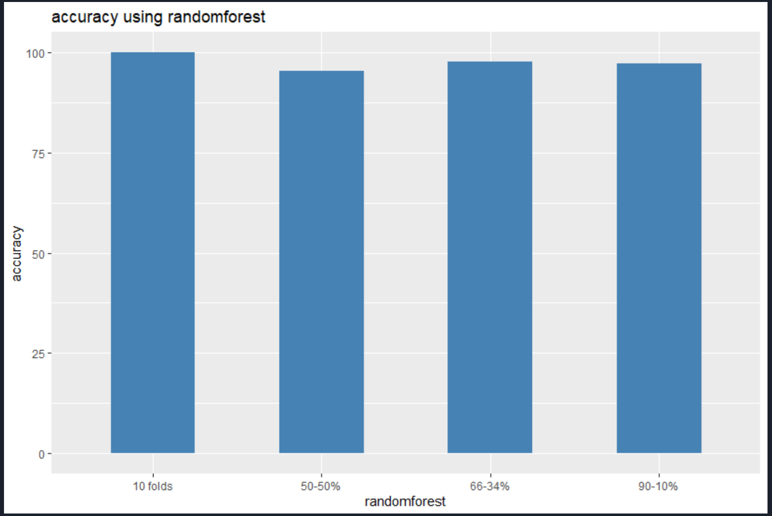
visualization of decision tree



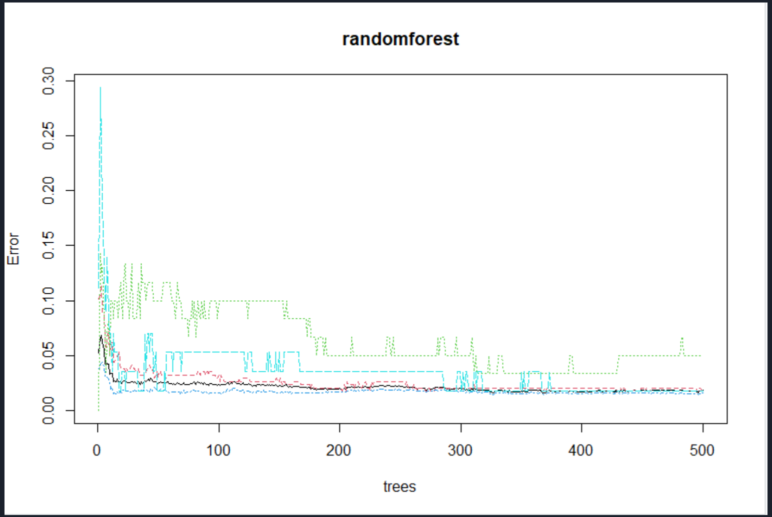
Randomforest[2]

- Random forest is composed of many decision trees(500 trees)
- A random forest has higher accuracy for data processing than a decision tree

Accuracy of random forest method



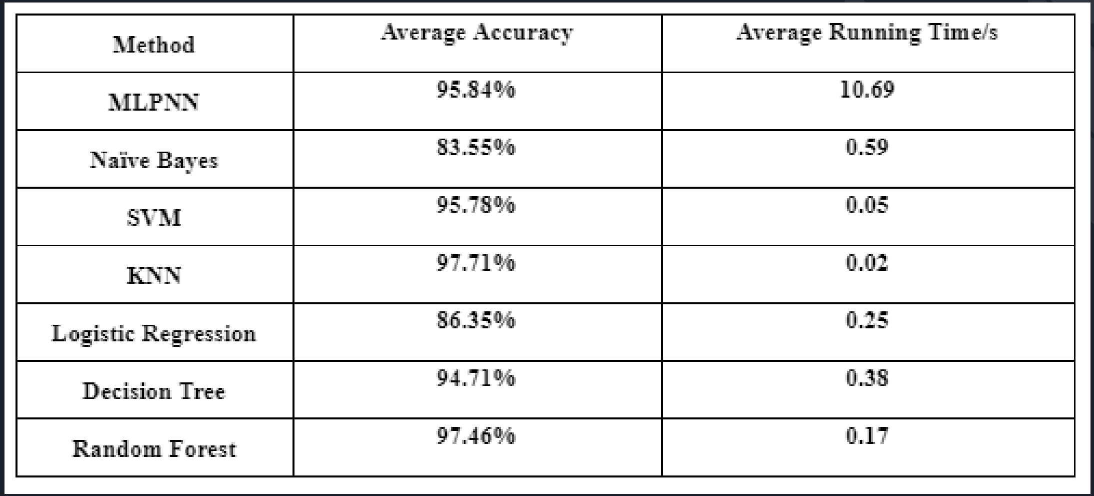
Randomforest 500 tree error model



Conclusion

- Only Naïve Bayes and Logistic Regression accuracy below 94%
- KNN performed the best whether in terms of accuracy and efficiency.
- Preprocess the data more suitable may help for the accuracy(LOOCV, Boosting and Adaboost)

Performance of each method



Method	Average Accuracy	Average Running Time/s
MLPNN	95.84%	10.69
Naïve Bayes	83.55%	0.59
SVM	95.78%	0.05
KNN	97.71%	0.02
Logistic Regression	86.35%	0.25
Decision Tree	94.71%	0.38
Random Forest	97.46%	0.17

Reference

[1] Gupta, P. (2017, November 12). Decision trees in machine learning. Medium. Retrieved December 19, 2021

[2] Maklin, C. (2019, July 30). Random Forest in R. Medium. Retrieved December 19, 2021

[3] Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. IEEE Intelligent Systems and their applications, 13(4), 18-28.

[4] Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003, November). KNN model-based approach in classification. In OTM Confederated International Conferences" On the Move to Meaningful Internet Systems" (pp. 986-996). Springer, Berlin, Heidelberg.

[5] de Souza, R. M., Cysneiros, F. J. A., Queiroz, D. C., & Roberta, A. D. A. (2008, October). A multi-class logistic regression model for interval data. In 2008 IEEE International Conference on Systems, Man and Cybernetics (pp. 1253-1258). IEEE.

Thanks for your listening!