# Part A Cross Validation

Cross-validation is prevalent in classification problems and used the data resampling method to assess the generalization of a predictor model in unseen data. In our case, determining suboptimal performance estimation is our primary goal, and the size of our dataset is considered when choosing a model validation. Decision tree is selected as our model because the Triage dataset is a multi-classification problem. To make the comparison square and fair, the hyperparameter of the decision tree is pre-set in constant where the complexity parameter is fixed at 0.001, min\_bucket, and pruning is set at 5. k-Fold Cross Validation and Leave-One-Out Cross Validation are compared and evaluated thru performance estimation based on the nature of our dataset Triage. Both cross validation techniques employ a different approach, with the Leave one out cross validation is bluntly straightforward by considering each of the observations as the validation set in the sense that it trains and validates the model for all possible combinations while K-fold cross validation is only considering the number of splits in the dataset by pre-specifying the K-value, for example, k=5, the dataset will be split into 5 groups of equal proportion of size, the first group is treated as a validation, and fitted on the remaining k-1 groups.

**Methodology**

*#Triage is a multi-classification dataset with inscript 5 factors; thus decision tree ML is deployed.*

*>model <- rpart(triage~., cp=0.001,maxdepth=5,minbucket=5,method='class',data=triageData)*

*# cp - complexity parameter*

*# maxdepth - max tree depth*

*# minbucket - min number of obs in leaf nodes*

*# method - return classification*

*options(repr.plot.width = 6, repr.plot.height = 6)*

*prp(model, space=4,split.cex=1.2,nn.border.col=0)*

*#Make prediction onto the dataset*

*train\_preds <- predict(model, newdata=triageData, type="class")*

*#Evaluation metric*

*confusionMatrix(factor(train\_preds), factor(triageData$triage))*

*#Cross-validation [k-fold CV vs LOOCV]*

*#k-fold CV=10*

*train\_control\_cv<- trainControl(method="cv",number=10)*

*#LOOCV*

*train\_control\_loocv <- trainControl(method = "LOOCV")*

*#set required parametes for the model type*

*tune\_grid = expand.grid(cp=c(0.001))*

*# Use the train() function to create the model*

*validated\_tree\_cv <- train(triage~.,method='rpart',data=triageData,maxdepth=5,minbucket=5, trControl=train\_control\_cv,tuneGrid=tune\_grid)*

*validated\_tree\_loocv <- train(triage~.,method='rpart',data=triageData,maxdepth=5,minbucket=5, trControl=train\_control\_loocv,tuneGrid=tune\_grid)*

*#summary of the model*

*validated\_tree\_cv*

*validated\_tree\_loocv*

**Model Evaluation**

Performance estimation of both cross-validation techniques is tabulated in table 1. As our dataset is a multi-classification problem set, thus it is evaluated in a contingency way.

Table 1.0: performance estimation of K-Fold Cross Validation and LOOCV.

|  |  |  |
| --- | --- | --- |
|  | 10-Fold Cross Validation | Leave-One-Out Cross Validation |
| Accuracy | 0.9987085 | 0.9961208 |
| Kappa | 0.9985232 | 0.9955609 |

Accuracy and Kappa are used to assess the inter-rater reliability of agreement. The Kappa coefficient indicates the interest in the magnitude of Kappa, values greater than 0.75 or so may be taken to present excellent agreement beyond chance. Simply put, the evaluation is deemed almost perfect for our model. While comparing 2 crossing over approaches, k-Fold Cross Validation and LOOCV, the difference between accuracy and kappa is minimal, which is less than 0.001. The advantage of LOOCV is a less biased model as almost every data point is used for training, wherein training takes place in the model on N-1 data points and testing the model against that one data point which was left in the previous step until all the data points are covered while its disadvantage is its expensive computation time. In another way, k-Fold Cross Validation offers a faster computation speed wherein the dataset has been split randomly into 10 subsets, one subset is used as a validation set and the remnants are involved in training the model. The disadvantage of k-Fold Cross Validation has been hedged by pre-specifying 10-folds as 10 is the most desired k value in most of the studies [1]. Thus, the 10-Fold Cross Validation technique is opted for our small dataset because of speedy computation time, promising performance estimation, and low bias a modest variance.

# 2.0 Part B Unsupervised ML

References:

[1] I. K. Nti, O. Nyarko-Boateng, and J. Aning, "Performance of Machine Learning Algorithms with Different K Values in K-fold CrossValidation," *International Journal of Information Technology and Computer Science,* vol. 13, no. 6, pp. 61-71, 2021.