####reference

# <https://cran.r-project.org/web/packages/DataExplorer/vignettes/dataexplorer-intro.html>

# <https://www.kaggle.com/sanchitakarmakar/mushroom-classification-99-75-ac-rf-pca>

# <https://www.kaggle.com/andreshg/mushroom-s-h2o-automl-and-clasic-models-auc-1-0>

# <https://www.kaggle.com/tosinabase/mushroom-classification-tree-methods-comparison>

# <https://www.kaggle.com/taimurzahid/mushroom-classification>

# <http://www.sthda.com/english/wiki/correlation-matrix-a-quick-start-guide-to-analyze-format-and-visualize-a-correlation-matrix-using-r-software>

# <https://www.displayr.com/how-to-create-a-correlation-matrix-in-r/>

# <https://stackoverflow.com/questions/27528907/how-to-convert-data-frame-column-from-factor-to-numeric/27528953>

# <https://towardsdatascience.com/how-to-create-a-correlation-matrix-with-too-many-variables-309cc0c0a57>

#####

# <https://www.machinelearningplus.com/machine-learning/caret-package/>

# <http://topepo.github.io/caret/available-models.html>

# <http://topepo.github.io/caret/train-models-by-tag.html>

#mehtods glm, naive bayes -> qda (one specific of naive bayes) -> lda (simplification of qda)

#predict(train\_qda, test\_set, type = "prob") %>% head()

#predict(train\_qda, test\_set) %>% head()

# rpart to create regression tree, can show descision flow chart

# train\_rpart <- train(margin ~ ., method = "rpart", tuneGrid = data.frame(cp = seq(0, 0.05, len = 25)), data = polls\_2008)

# can tune CP here

# Classification trees, or decision trees, are used in prediction problems where the outcome is categorical.

# Decision trees form predictions by calculating which class is the most common among the training set observations

# within the partition, rather than taking the average in each partition.

# Then can introduce random forest

# Random Forest

# fit <- train(y ~ ., method = "rRborist", tuneGrid = data.frame, data = )

# use mtry to select the features used in prediction by random

# use variable importance to see which variable is important

# carat package reference

# http://topepo.github.io/caret/available-models.html

# http://topepo.github.io/caret/train-models-by-tag.html

# y\_hat\_glm <- predict(train\_glm, mnist\_27$test, type = "raw")

# y\_hat\_knn <- predict(train\_knn, mnist\_27$test, type = "raw")

# ggplot(train\_knn\_cv, highlight = TRUE)

# train can do the parameter maximization

# ggplot(train\_knn, highlight = TRUE)

# train\_knn$bestTune

# train\_knn$finalModel

# can use change control

#control <- trainControl(method = "cv", number = 10, p = .9)

#train\_knn\_cv <- train(y ~ ., method = "knn",

# data = mnist\_27$train,

# tuneGrid = data.frame(k = seq(9, 71, 2)),

#ggplot(train\_knn\_cv, highlight = TRUE)

# train\_knn$results %>%

# ggplot(aes(x = k, y = Accuracy)) +

# geom\_line() +

# geom\_point() +

# geom\_errorbar(aes(x = k,

# ymin = Accuracy - AccuracySD,

# ymax = Accuracy + AccuracySD)) trControl = control)

# knn then can use gamLoess

# pre-processing to transform predictr and remove predictor not useful

# are highly correlated with others, have very few non-unique values, or have close to zero variation

# caret package : nzv <- nearZeroVar(x)

# caret need to have column name

# variable importance importance()

# ensemble different model to one

**Attribute Information:**

1. cap-shape: bell=b,conical=c,convex=x,flat=f, knobbed=k,sunken=s  
2. cap-surface: fibrous=f,grooves=g,scaly=y,smooth=s  
3. cap-color: brown=n,buff=b,cinnamon=c,gray=g,green=r, pink=p,purple=u,red=e,white=w,yellow=y  
4. bruises?: bruises=t,no=f  
5. odor: almond=a,anise=l,creosote=c,fishy=y,foul=f, musty=m,none=n,pungent=p,spicy=s  
6. gill-attachment: attached=a,descending=d,free=f,notched=n  
7. gill-spacing: close=c,crowded=w,distant=d  
8. gill-size: broad=b,narrow=n  
9. gill-color: black=k,brown=n,buff=b,chocolate=h,gray=g, green=r,orange=o,pink=p,purple=u,red=e, white=w,yellow=y  
10. stalk-shape: enlarging=e,tapering=t  
11. stalk-root: bulbous=b,club=c,cup=u,equal=e, rhizomorphs=z,rooted=r,missing=?  
12. stalk-surface-above-ring: fibrous=f,scaly=y,silky=k,smooth=s  
13. stalk-surface-below-ring: fibrous=f,scaly=y,silky=k,smooth=s  
14. stalk-color-above-ring: brown=n,buff=b,cinnamon=c,gray=g,orange=o, pink=p,red=e,white=w,yellow=y  
15. stalk-color-below-ring: brown=n,buff=b,cinnamon=c,gray=g,orange=o, pink=p,red=e,white=w,yellow=y  
16. veil-type: partial=p,universal=u  
17. veil-color: brown=n,orange=o,white=w,yellow=y  
18. ring-number: none=n,one=o,two=t  
19. ring-type: cobwebby=c,evanescent=e,flaring=f,large=l, none=n,pendant=p,sheathing=s,zone=z  
20. spore-print-color: black=k,brown=n,buff=b,chocolate=h,green=r, orange=o,purple=u,white=w,yellow=y  
21. population: abundant=a,clustered=c,numerous=n, scattered=s,several=v,solitary=y  
22. habitat: grasses=g,leaves=l,meadows=m,paths=p, urban=u,waste=w,woods=d