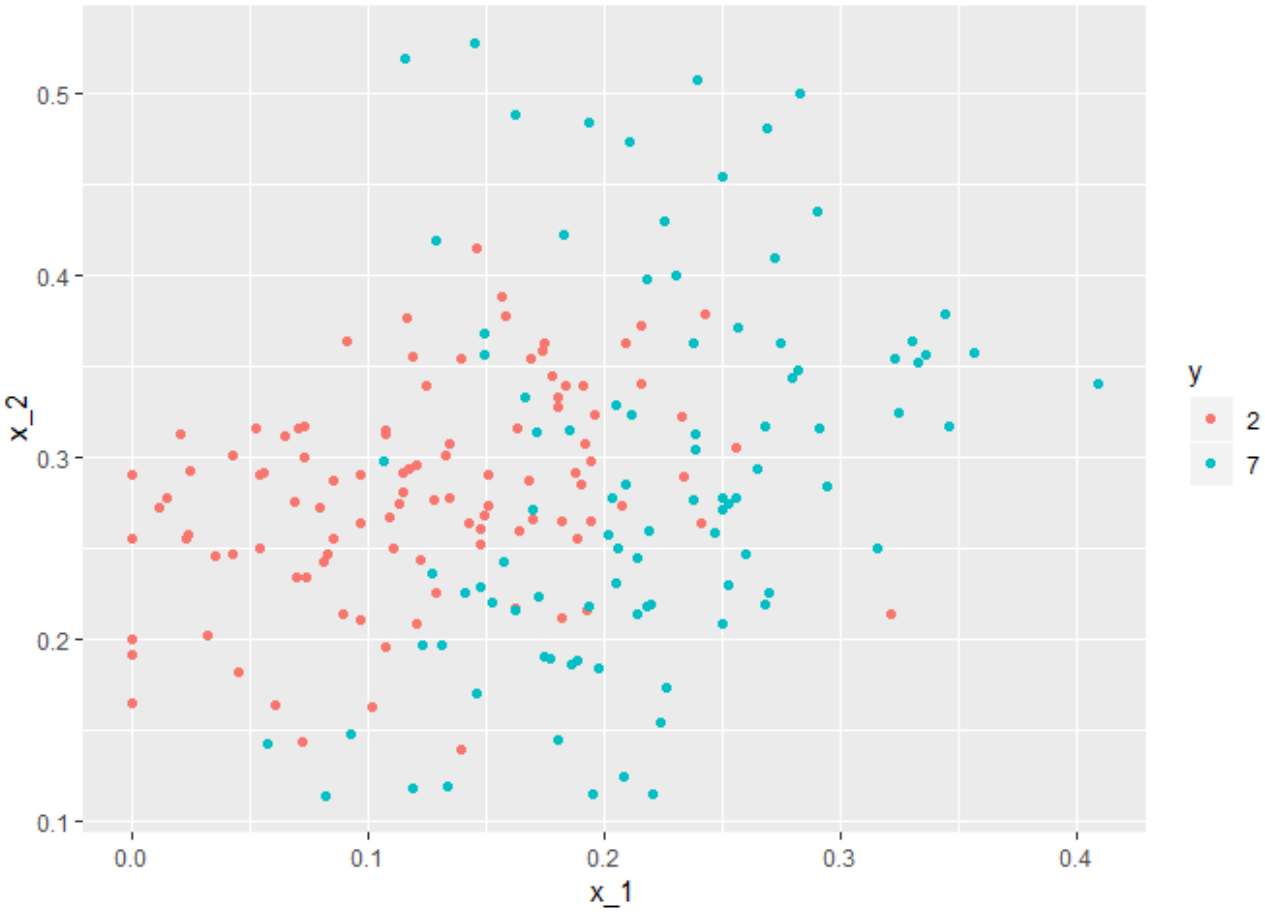
MNIST

MNIST prediction with KNN

Load the data

```
data("mnist_27")
# Plot the data for digits 2 and 7
mnist_27$test%>% ggplot(aes(x_1, x_2, color = y)) + geom_point()
```



Create the train set

```
x <- as.matrix(mnist_27$train[,2:3])
y <- mnist_27$train$y
```

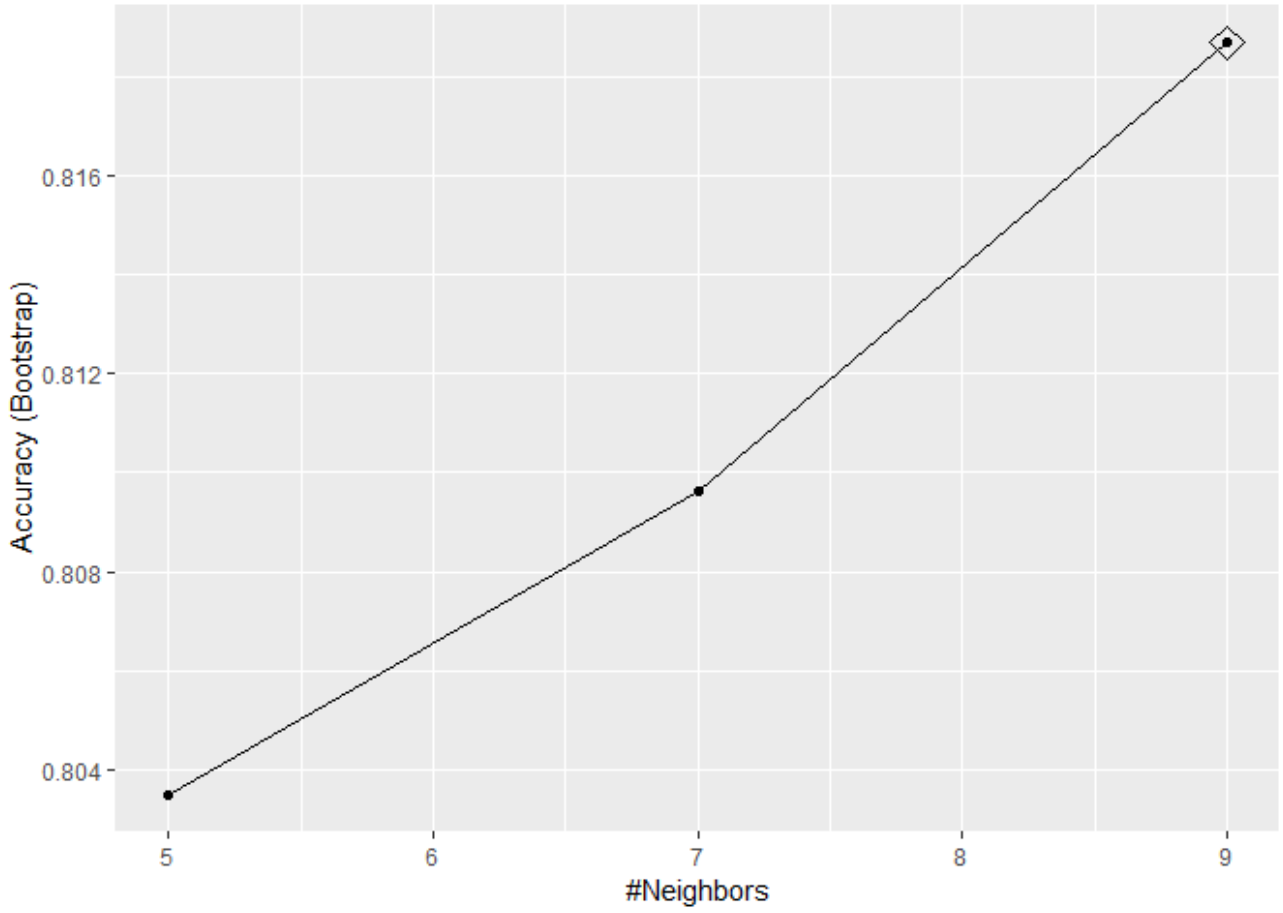
Modeling

```
knn_fit <- knn3(y ~ ., data = mnist_27$train, k = 5)
# Prediction
y_hat_knn <- predict(knn_fit, mnist_27$test, type= "class")
# Confusion Marix
confusionMatrix(data= y_hat_knn, reference = mnist_27$test$y)$overall["Accuracy"]
```

```
## Accuracy
##    0.815
```

Crossvalidation with default values in caret

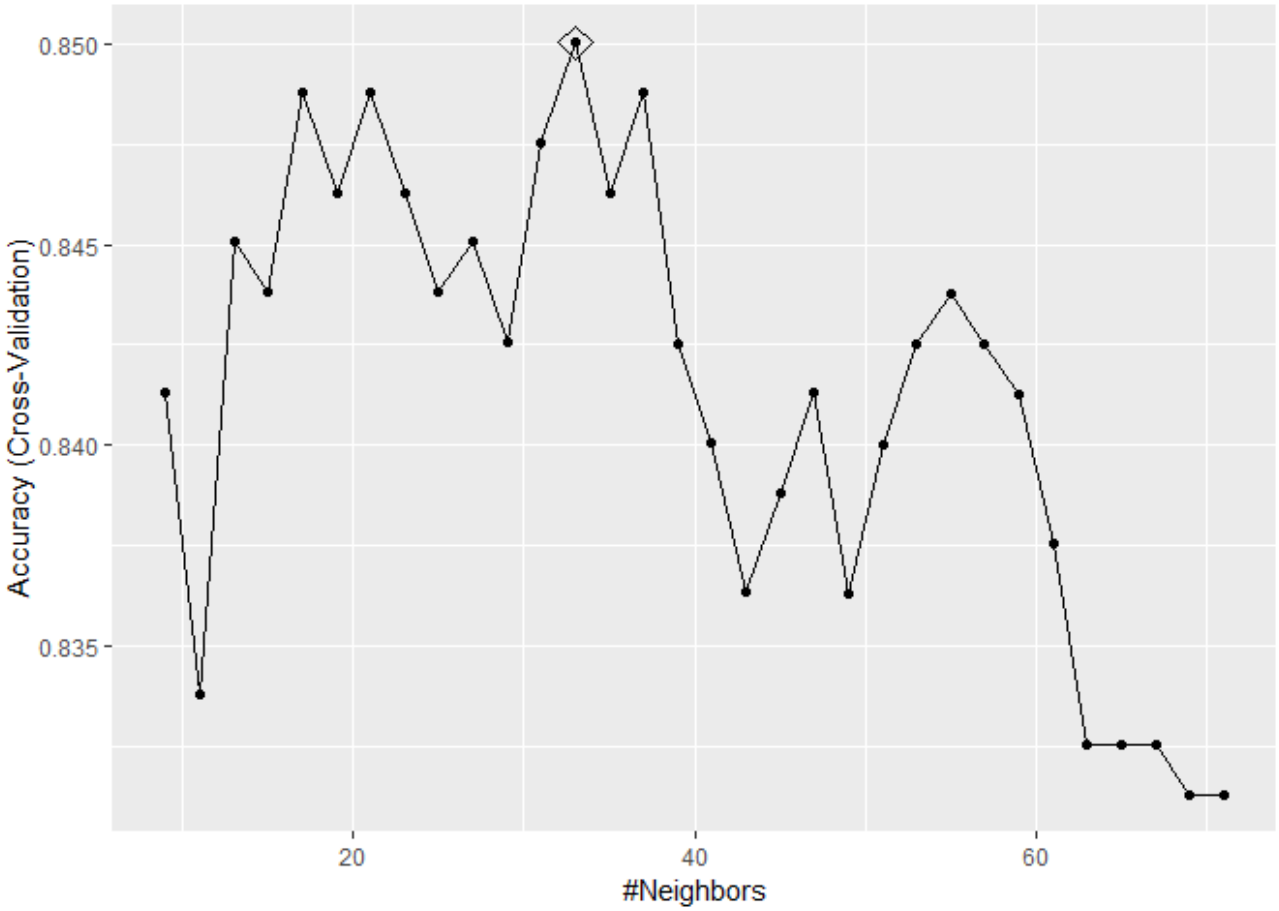
```
train_knn <- train(y ~ ., method = "knn", data = mnist_27$train)
ggplot(train_knn, highlight = TRUE)
```



Crossvalidation with trainControl and tuneGrid

```
control <- trainControl(method = "cv", number = 10, p = 0.9)
train_knn_cv <- train(y ~ ., method = "knn",
                      data = mnist_27$train,
                      tuneGrid = data.frame(k = seq(9, 71, 2)),
```

```
                       trControl = control)
ggplot(train_knn_cv, highlight = TRUE)
```




```
train_knn_cv$bestTune
```

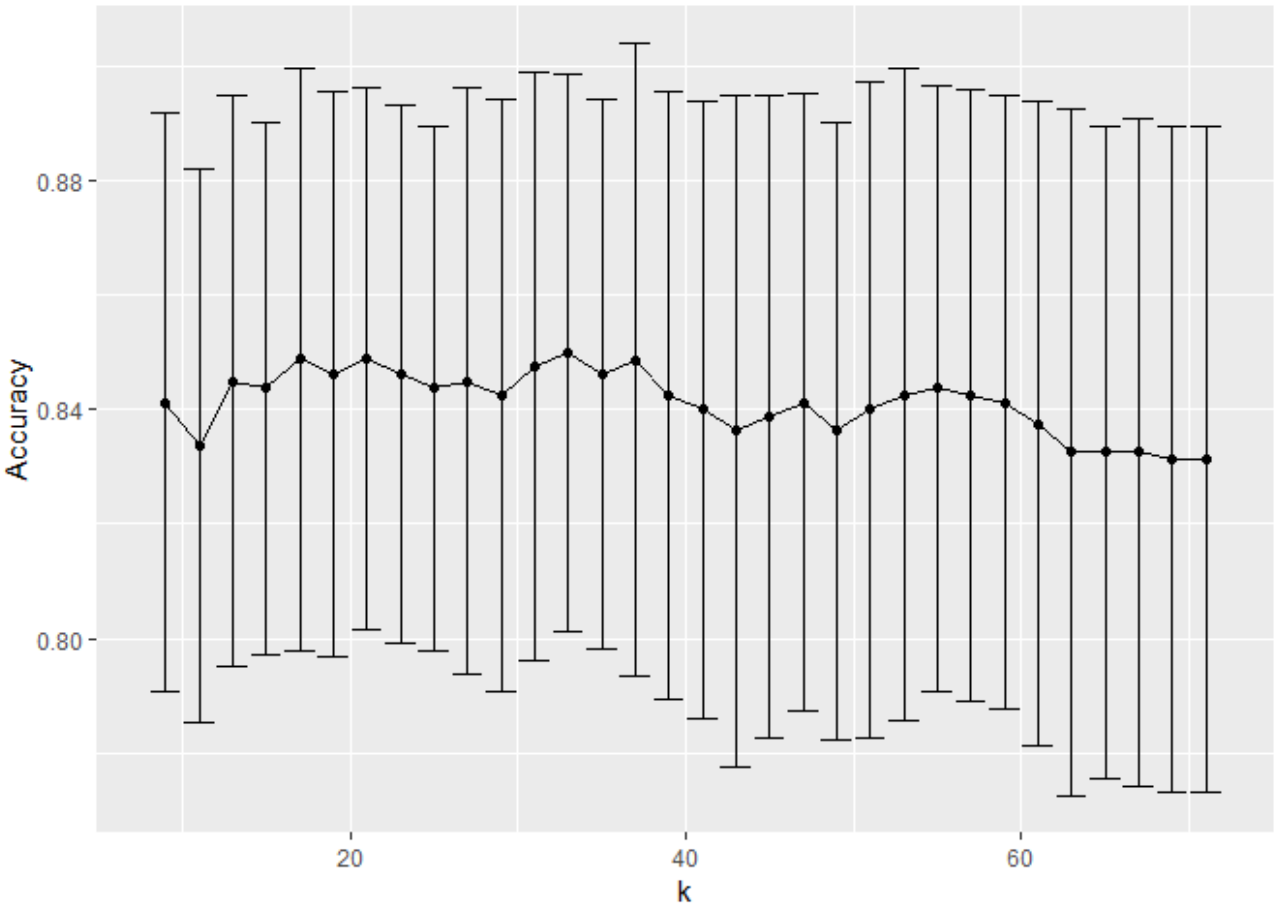
```
##     k
## 13 33
```

```
train_knn_cv$finalModel
```

```
## 33-nearest neighbor model
## Training set outcome distribution:
##
##   2   7
## 379 421
```

Plot the accuracy

```
train_knn_cv$results %>%
  ggplot(aes(x = k, y = Accuracy)) +   geom_line() +   geom_point() +
  geom_errorbar(aes(x = k,
                    ymin = Accuracy - AccuracySD,
                    ymax = Accuracy + AccuracySD))
```



Prediction

```
Pred_knn_cv <- predict(train_knn_cv, mnist_27$test,type="raw")
# confusion matrix
cm <- confusionMatrix(Pred_knn_cv, mnist_27$test$y)
cm
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  2  7
##          2 93 18
##          7 13 76
##
##                Accuracy : 0.845
##                  95% CI : (0.7873, 0.8922)
```

```
##     No Information Rate : 0.53
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.6879
##  Mcnemar's Test P-Value : 0.4725
##
##             Sensitivity : 0.8774
##             Specificity : 0.8085
##          Pos Pred Value : 0.8378
##          Neg Pred Value : 0.8539
##              Prevalence : 0.5300
##          Detection Rate : 0.4650
##    Detection Prevalence : 0.5550
##       Balanced Accuracy : 0.8429
##
##        'Positive' Class : 2
##
```

```
cm$overall["Accuracy"]
```

```
## Accuracy
##    0.845
```