

# Breast Cancer

*Shashank Semwal*

*25 June 2018*

## INSERT LIBRARIES

```
library(class)
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.4.4
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

### Import DATA

```
wdbc<-read.table("C:/Users/Shashank/Documents/R/dataset/breast cancer/wdbc.data",sep=',')
dim(wdbc)
```

```
## [1] 569 32
```

### Removing the lables of the data & creating samples

```
wdbc_sample=sample(nrow(wdbc),size = nrow(wdbc)*.7)
wdbc_train=wdbc[wdbc_sample,-c(1,2)]
wdbc_test=wdbc[-wdbc_sample,-c(1,2)]
```

### Standarizing the data

```
wdbc_std_train<-as.data.frame(lapply(wdbc_train,function(x) (x-min(x))/(max(x)-min(x))))  
wdbc_std_test<-as.data.frame(lapply(wdbc_test,function(x) (x-min(x))/(max(x)-min(x))))  
dim(wdbc_std_test)
```

```
## [1] 171 30
```

```
predict<-knn(train =wdbc_std_train,test = wdbc_std_test,cl = wdbc[wdbc_sample,2],k = 3 )
```

```
confusionMatrix(predict,wdbc[-wdbc_sample,2])
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  B    M
##           B 106   1
##           M   9   55
##
##           Accuracy : 0.9415
##           95% CI : (0.8951, 0.9716)
##           No Information Rate : 0.6725
##           P-Value [Acc > NIR] : < 2e-16
##
##           Kappa : 0.8719
##           McNemar's Test P-Value : 0.02686
##
##           Sensitivity : 0.9217
##           Specificity : 0.9821
##           Pos Pred Value : 0.9907
##           Neg Pred Value : 0.8594
##           Prevalence : 0.6725
##           Detection Rate : 0.6199
##           Detection Prevalence : 0.6257
##           Balanced Accuracy : 0.9519
##
##           'Positive' Class : B
##
```

```
ggplot(,aes(wdbc[-wdbc_sample,3],wdbc[-wdbc_sample,4],col=predict))+
  geom_jitter(stat = 'identity')
```

