**Task 1: Import the dataset**

**library**(readr)

Zombie\_Dataset <- **read\_csv**("C:/Users/tra/Desktop/assignment 2/Zombie Dataset. csv",

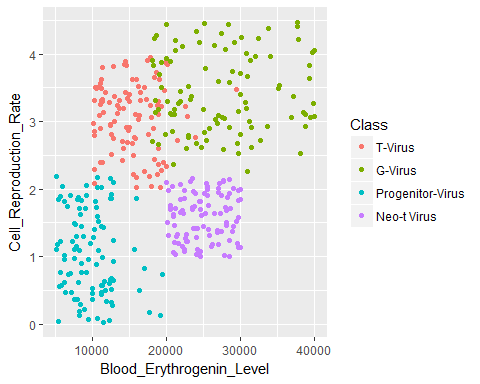
col\_types = **cols**(Class = **col\_factor**(levels = **c**("T-Virus", "G-Virus", "Progenitor-Virus", "Neo-t Virus"))))

mydata=Zombie\_Dataset

**Task 2: Visualize the data**

• From plot T-Virus and G-Virus seem to be mixed together in some section **library**(ggplot2) **qplot**(Blood\_Erythrogenin\_Level,Cell\_Reproduction\_Rate,data=mydata,color=Class

)



**Task 3: Default SVM**

• Number of support vectors are 125

**library**(e1071)

mymodel <- **svm**(Class~.,data=mydata)

**summary**(mymodel)

##

## Call:

## svm(formula = Class ~ ., data = mydata)

##

##

## Parameters:

## SVM-Type: C-classification

## SVM-Kernel: radial

## cost: 1

## gamma: 0.5

##

## Number of Support Vectors: 125

##

## ( 45 41 16 23 )

##

##

## Number of Classes: 4

##

## Levels:

## T-Virus G-Virus Progenitor-Virus Neo-t Virus

**Task 4: Prediction and Evaluation**

• The accuracy is 94.25%

pred<-**predict**(mymodel,mydata)

performance<-**table**(predicted=pred,actual=mydata$Class)

performance

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ## |  | actual |  | | | |
| ## | predicted | T-Virus | G-Virus | Progenitor-Virus | Neo-t | Virus |
| ## | T-Virus | 87 | 7 | 1 |  | 1 |
| ## | G-Virus | 11 | 92 | 0 |  | 0 |
| ## | Progenitor-Virus | 2 | 0 | 99 |  | 0 |
| ## | Neo-t Virus | 0 | 1 | 0 |  | 99 |

accuracy=**sum**(**diag**(performance)/**sum**(performance))

accuracy

## [1] 0.9425

**Task 5: Tune the Model**

• Yes the best model differs from the default model

• The cost factor is to assign pentaly to the misclassified points. So if the cost parameter increases that means we will have less misclassified points.

• Now we have 85 supporting vectors in the best model

• The accuracy of the best model is 94.75%. The best model has better accuracy than the default model

**set.seed**(123)

tunedmodel<- **tune**(svm,Class~.,data=mydata, ranges=**list**(cost=**c**(0.1, 1, 10), ga mma=**c**(0.25, 0.5, 1, 4), kernel="radial"))

**summary**(tunedmodel)

##

## Parameter tuning of 'svm':

##

## - sampling method: 10-fold cross validation

##

## - best parameters:

## cost gamma kernel

## 10 0.25 radial

##

## - best performance: 0.06

##

## - Detailed performance results:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ## |  | cost | gamma | kernel | error | dispersion |
| ## | 1 | 0.1 | 0.25 | radial | 0.0900 | 0.04116363 |
| ## | 2 | 1.0 | 0.25 | radial | 0.0700 | 0.02838231 |
| ## | 3 | 10.0 | 0.25 | radial | 0.0600 | 0.03574602 |
| ## | 4 | 0.1 | 0.50 | radial | 0.0800 | 0.03496029 |
| ## | 5 | 1.0 | 0.50 | radial | 0.0700 | 0.03496029 |
| ## | 6 | 10.0 | 0.50 | radial | 0.0625 | 0.03385016 |
| ## | 7 | 0.1 | 1.00 | radial | 0.0725 | 0.02993047 |
| ## | 8 | 1.0 | 1.00 | radial | 0.0650 | 0.03374743 |
| ## | 9 | 10.0 | 1.00 | radial | 0.0675 | 0.04257347 |
| ## | 10 | 0.1 | 4.00 | radial | 0.0700 | 0.04048319 |
| ## | 11 | 1.0 | 4.00 | radial | 0.0675 | 0.03736085 |
| ## | 12 | 10.0 | 4.00 | radial | 0.0700 | 0.04377975 |

svm\_model\_after\_tune <- **svm**(Class ~ ., data=mydata, kernel="radial", cost=10, gamma=0.25)

**summary**(svm\_model\_after\_tune)

##

## Call:

## svm(formula = Class ~ ., data = mydata, kernel = "radial", cost = 10,

## gamma = 0.25)

##

##

## Parameters:

## SVM-Type: C-classification

## SVM-Kernel: radial

## cost: 10

## gamma: 0.25

##

## Number of Support Vectors: 85

##

## ( 33 30 11 11 )

##

##

## Number of Classes: 4

##

## Levels:

## T-Virus G-Virus Progenitor-Virus Neo-t Virus

pred1<-**predict**(svm\_model\_after\_tune,mydata)

performance1<-**table**(predicted=pred1,actual=mydata$Class)

performance1

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ## |  | actual |  | | | |
| ## | predicted | T-Virus | G-Virus | Progenitor-Virus | Neo-t | Virus |
| ## | T-Virus | 87 | 7 | 0 |  | 1 |
| ## | G-Virus | 11 | 93 | 0 |  | 0 |
| ## | Progenitor-Virus | 2 | 0 | 100 |  | 0 |
| ## | Neo-t Virus | 0 | 0 | 0 |  | 99 |

accuracy1=**sum**(**diag**(performance1)/**sum**(performance1))

accuracy1

## [1] 0.9475

**Task 6: Prepare Data for Clustering**

mydata1= mydata[-3]

scaleddata=**scale**(mydata1)

**Task 7: Hierarchical Clustering**

pairwisedistance=**dist**(scaleddata)

hc.c=**hclust**(pairwisedistance)

**plot**(hc.c)

