Question-2.

a) Now consider, from the KC weather data set, just the predictors: Temp.F, Humidity. Percentage, Precip.in. Categorize these three data sets into qualitative predictors. It is up to you to decide on the break points, but you must discuss a rationale for your breakpoints. Now apply, naive Bayes Classifier on the entire data set (with these three qualitative predictors), using 290 of them as a training data set randomly (and the rest as the test data set), over 100 replications. Report on accuracy, precision, and recall.

A.

- **Qualitative Predictors**

```
require(caret)
require(tm)
require(wordcloud)
require(e1071)
require(MLmetrics)
library('e1071')
weather1<-read.csv(file="C:\\Users\\sudheesha\\Downloads\\kc_weather_srt.csv")
Data2=weather1[,c("Temp.F","Humidity.percentage","Precip.in","Events")]
head(Data2)
Data2$Temp.F[Data2$Temp.F < 10] <- 'T_1s'
Data2$Temp.F[Data2$Temp.F >=10 & Data2$Temp.F <20] <- 'T_10s'
Data2$Temp.F[Data2$Temp.F >= 20 & Data2$Temp.F <30] <- 'T_20s'
Data2$Temp.F[Data2$Temp.F >= 30 & Data2$Temp.F <40 ] <- 'T_30s'
Data2$Temp.F[Data2$Temp.F >= 40 & Data2$Temp.F <50 ] <- 'T_40s'
Data2$Temp.F[Data2$Temp.F >= 50 & Data2$Temp.F <60 ] <- 'T_50s'
Data2$Temp.F[Data2$Temp.F >= 60 & Data2$Temp.F <70 ] <- 'T_60s'
Data2$Temp.F[Data2$Temp.F >= 70 & Data2$Temp.F <80 ] <- 'T_70s'
Data2$Temp.F[Data2$Temp.F >= 80 & Data2$Temp.F <90 ] <- 'T_80s'
Data2$Humidity.percentage[Data2$Humidity.percentage>=20       &
Data2$Humidity.percentage <40] <- 'H_20s_30s'
Data2$Humidity.percentage[Data2$Humidity.percentage>=40       &
Data2$Humidity.percentage <50 ] <- 'H_40s'
Data2$Humidity.percentage[Data2$Humidity.percentage>=50       &
Data2$Humidity.percentage <70 ] <- 'H_50s_60s'
Data2$Humidity.percentage[Data2$Humidity.percentage>=70       &
Data2$Humidity.percentage <90 ] <- 'H_70s_80s'
Data2$Humidity.percentage[Data2$Humidity.percentage>=90       &
Data2$Humidity.percentage <99 ] <- 'H_90s'
Data2$Precip.in[Data2$Precip.in == 0] <- 'P_0s'
Data2$Precip.in[Data2$Precip.in >0 & Data2$Precip.in < 1] <- 'P_0.01s'
Data2$Precip.in[Data2$Precip.in >=2 & Data2$Precip.in <3 ] <- 'P_2s'
entries=366
training=290
test=entries-training
```

```
replica=100
accuracy=dim(replica)
precision=dim(replica)
recall=dim(replica)
elinear=dim(replica)
for (i in 1:replica)
{
  train=sample(1:entries,training)
  weather2.nb = naiveBayes(Events~.,Data2[train,])
  Data2.test = Data2[-train,1:3]
  predict(weather2.nb, Data2.test, type="raw")
  tablin=table(predict(kcweather.nb,Data2.test,type="class"),Data2[-train,4])
  accuracy[i] = (tablin[1,1]+tablin[2,2])/(sum(tablin))
  precision[i] = (tablin[1,1])/(tablin[1,1]+tablin[2,1])
  recall[i]=(tablin[1,1])/(tablin[1,1]+tablin[1,2])
}
cat("Accuracy: ",mean(accuracy))
cat("Precision :  ",mean(precision))
cat("Recall: ",mean(recall))
```

Output:

```
> cat("Accuracy: ",mean(accuracy))
Accuracy:  0.4853947> cat("Precision :  ",mean(precision))
Precision :   0.9930711> cat("Recall: ",mean(recall))
Recall:  0.5628827>
```

- **Temperature as quantitave predictor**

```
require(caret)
require(tm)
require(wordcloud)
require(e1071)
require(MLmetrics)
library('e1071')
weather2<-
read.csv(file="C:\\Users\\sudheesha\\Downloads\\kc_weather_srt.csv")
Data2=weather2[,c("Temp.F","Humidity.percentage","Precip.in","Events")]
Data2$Humidity.percentage[Data2$Humidity.percentage>=20       &
Data2$Humidity.percentage <40] <- 'H_20s_30s'
Data2$Humidity.percentage[Data2$Humidity.percentage>=40       &
Data2$Humidity.percentage <50 ] <- 'H_40s'
Data2$Humidity.percentage[Data2$Humidity.percentage>=50       &
Data2$Humidity.percentage <70 ] <- 'H_50s_60s'
```

```
Data2$Humidity.percentage[Data2$Humidity.percentage>=70       &
Data2$Humidity.percentage <90 ] <- 'H_70s_80s'
Data2$Humidity.percentage[Data2$Humidity.percentage>=90       &
Data2$Humidity.percentage <99 ] <- 'H_90s'
Data2$Precip.in[Data2$Precip.in == 0] <- 'P_0s'
Data2$Precip.in[Data2$Precip.in >0 & Data2$Precip.in < 1] <- 'P_0.01s'
Data2$Precip.in[Data2$Precip.in >=2 & Data2$Precip.in <3 ] <- 'P_2s'
entries=366
training=290
test=entries-training
replica=100
accuracy=dim(replica)
precision=dim(replica)
recall=dim(replica)
elinear=dim(replica)
for (i in 1:replica)
{
  train=sample(1:entries,training)
  .nb = naiveBayes(Events~.,Data2[train,])
  Data2.test = Data2[-train,1:3]
  predict(weather2.nb, Data2.test, type="raw")
  tablin=table(predict(weather2.nb,Data2.test,type="class"),Data2[-train,4])
  accuracy[i] = (tablin[1,1]+tablin[2,2])/(sum(tablin))
  precision[i] = (tablin[1,1])/(tablin[1,1]+tablin[2,1])
  recall[i]=(tablin[1,1])/(tablin[1,1]+tablin[1,2])
}
cat("Accuracy: ",mean(accuracy))
cat("Precision :  ",mean(precision))
cat("Recall: ",mean(recall))
```

Output

```
> cat("Accuracy: ",mean(accuracy))
Accuracy:  0.4769737> cat("Precision :  ",mean(precision))
Precision :   0.9898143> cat("Recall: ",mean(recall))
Recall:  0.5530994>
```

- **All predictors as quantitative predictors**

```
require(caret)

require(tm)

require(wordcloud)

require(e1071)
```

```
require(MLmetrics)

library('e1071')

weather2<-
read.csv(file="C:\\Users\\sudheesha\\Downloads\\kc_weather_srt.csv")

Data2=weather2[,c("Temp.F","Humidity.percentage","Precip.in","Events")]

entries=366
training=290
test=entries-training
replica=100
accuracy=dim(replica)
precision=dim(replica)
recall=dim(replica)
elinear=dim(replica)
for (i in 1:replica)

{

  train=sample(1:entries,training)

  weather2.nb = naiveBayes(Events~.,Data2[train,])

  Data2.test = Data2[-train,1:3]

  predict(weather2.nb, Data2.test, type="raw")

  tablin=table(predict(weather2.nb,Data2.test,type="class"),Data2[-train,4])

  print(tablin)

  accuracy[k] = (tablin[1,1]+tablin[2,2])/(sum(tablin))

  precision[k] = (tablin[1,1])/(tablin[1,1]+tablin[2,1])

  recall[k]=(tablin[1,1])/(tablin[1,1]+tablin[1,2])

}

cat("Accuracy: ",mean(accuracy))

cat("Precision :  ",mean(precision))

cat("Recall: ",mean(recall))
```

Output:

```
> cat("Accuracy: ",mean(accuracy))
Accuracy:  0.6148684> cat("Precision :  ",mean(precision))
Precision :   0.787583> cat("Recall: ",mean(recall))
Recall:  0.7561261>
>
```

Summary

Naïve Bayes	Accuracy	Precision	Recall
Three Qualitative	0.4853947	0.9930711	0.5628827
Temp Quantitative	0.4769737	0.9898143	0.5530994
Three Quantitative	0.6148684	0.787853	0.7561261

Question1 Summary

Summary:

Model	Accuracy	Precision	Recall
LDA	0.9238136	0.8292378	0.8236089
QDA	0.9284047	0.7934406	0.9310367
KNN(k=3)	0.5596053	0.677384	0.7164928
KNN(k=10)	0.5410526	0.6760336	0.6876788

Conclusion

- **Three Qualitative:** Highest Precision, when precision is major significant
- **Three Quantitative:** Highest Accuracy, when accuracy is major significant
- **Temp Quantitaive:** Average values in all the cases

So finally, when the precision is more important than the model should be chosen as three qualitative and when the accuracy is more significant than the model should choose all the predictors quantitative. To conclude QDA is the best model when compared to Naïve bayes and Naïve bayes is better when compared to that of KNN(K=3) and KNN(K=10)