Problem 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.
b)

Here we use Naïve Baye's Classification:

```
> library(e1071)
Warning message:
package 'e1071' was built under R version 3.4.2
> library('e1071')
> nbdata=data[,c("Temp.F","Humidity.percentage","Precip.in","Events")]
> head(nbdata)
  Temp.F Humidity.percentage Precip.in Events
1     26                  73      0.03   Snow
2     31                  68      0.01   Snow
3     10                  63      0.02   Snow
4     38                  90      0.00   Rain
5     40                  75      0.00   Rain
6     49                  51      0.00   Rain
> nbdata$Temp.F[nbdata$Temp.F < 10] <- 'T_1s'
> nbdata$Temp.F[nbdata$Temp.F >=10 & nbdata$Temp.F <20] <- 'T_10s'
> nbdata$Temp.F[nbdata$Temp.F >= 20 & nbdata$Temp.F <30] <- 'T_20s'
> nbdata$Temp.F[nbdata$Temp.F >= 30 & nbdata$Temp.F <40 ] <- 'T_30s'
> nbdata$Temp.F[nbdata$Temp.F >= 40 & nbdata$Temp.F <50 ] <- 'T_40s'
> nbdata$Temp.F[nbdata$Temp.F >= 50 & nbdata$Temp.F <60 ] <- 'T_50s'
> nbdata$Temp.F[nbdata$Temp.F >= 60 & nbdata$Temp.F <70 ] <- 'T_60s'
> nbdata$Temp.F[nbdata$Temp.F >= 70 & nbdata$Temp.F <80 ] <- 'T_70s'
> nbdata$Temp.F[nbdata$Temp.F >= 80 & nbdata$Temp.F <90 ] <- 'T_80s'
>
> nbdata$Humidity.percentage[nbdata$Humidity.percentage>=20& nbdata$Humidity.percentage <40] <- 'H_20s_30s'
> nbdata$Humidity.percentage[nbdata$Humidity.percentage>=40& nbdata$Humidity.percentage <50 ] <- 'H_40s'
> nbdata$Humidity.percentage[nbdata$Humidity.percentage>=50& nbdata$Humidity.percentage <70 ] <- 'H_50s_60s'
> nbdata$Humidity.percentage[nbdata$Humidity.percentage>=70& nbdata$Humidity.percentage <90 ] <- 'H_70s_80s'
> nbdata$Humidity.percentage[nbdata$Humidity.percentage>=90& nbdata$Humidity.percentage <99 ] <- 'H_90s'
>
```

```
>
> nbdata$Precip.in[nbdata$Precip.in == 0] <- 'P_0s'
> nbdata$Precip.in[nbdata$Precip.in >0 & nbdata$Precip.in < 1] <- 'P_0.01s'
> nbdata$Precip.in[nbdata$Precip.in >=2 & nbdata$Precip.in <3 ] <- 'P_2s'
>
>
> n=366
> nt=290
> neval=n-nt
> rep=100
> errlin=dim(rep)
> accuracy=dim(rep)
> precision=dim(rep)
> recall=dim(rep)
> for (k in 1:rep)
+ {
+   train=sample(1:n,nt)
+   kcweather.nb = naiveBayes(Events~.,nbdata[train,])
+   nbdata.test = nbdata[-train,1:3]
+   predict(kcweather.nb, nbdata.test, type="raw")
+   tablin=table(predict(kcweather.nb,nbdata.test,type="class"),nbdata[-train,4])
+   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])
+ }
There were 50 or more warnings (use warnings() to see the first 50)
> cat("Accuracy: ",mean(accuracy))
Accuracy:  0.4846053> cat("Precision :  ",mean(precision))
Precision :   0.9945185> cat("Recall: ",mean(recall))
Recall:  0.5633328> |
```

We could see the accuracy, precision and recall values.

Accuracy: 0.484605

Precision: 0.9945185

Recall: 0.5633328

The break points for conversion of Temp, humidity, precipitation to make it qualitative.

Temperature:

Temp < 10 – T_1’s, Temp >=10 and Temp < 20 – T_10s

Temp >=20 and Temp < 30 – T_20s, Temp >=30 and Temp < 40 – T_30s

Temp >=40 and Temp < 50 – T_40s, Temp >=50 and Temp < 60 – T_50s

Temp >=60 and Temp < 70 – T_60s, Temp >=70 and Temp < 80 – T_70s

Temp >=80 and Temp < 90 – T_80s

Humidity:

Humidity >=20 and Humidity < 40 - 'H_20s_30s', Humidity >=40 and Humidity < 50 - 'H_40s'

Humidity >=50 and Humidity < 70 - 'H_50s_60s'Humidity >=70 and Humidity < 90 - 'H_70s_80s'

Humidity >=90 and Humidity < 99 - 'H_90s'

Precipitation:

Precipitation == 0 - 'P_0s', Precipitation > 0 & Precipitation < 1 - ‘P_0.01s'

Precipitation >=2 & Precipitation <3   - 'P_2s'

_Now we shall use Temp as the only quantitative predictor:

```
> nbdata=data[,c("Temp.F","Humidity.percentage","Precip.in","Events")]
>
> nbdata$Humidity.percentage[nbdata$Humidity.percentage>=20& nbdata$Humidity.percentage <40] <- 'H_20s_30s'
> nbdata$Humidity.percentage[nbdata$Humidity.percentage>=40& nbdata$Humidity.percentage <50 ] <- 'H_40s'
> nbdata$Humidity.percentage[nbdata$Humidity.percentage>=50& nbdata$Humidity.percentage <70 ] <- 'H_50s_60s
> nbdata$Humidity.percentage[nbdata$Humidity.percentage>=70& nbdata$Humidity.percentage <90 ] <- 'H_70s_80s
> nbdata$Humidity.percentage[nbdata$Humidity.percentage>=90& nbdata$Humidity.percentage <99 ] <- 'H_90s'
> nbdata$Precip.in[nbdata$Precip.in == 0] <- 'P_0s'
> nbdata$Precip.in[nbdata$Precip.in >0 & nbdata$Precip.in < 1] <- 'P_0.01s'
> nbdata$Precip.in[nbdata$Precip.in >=2 & nbdata$Precip.in <3 ] <- 'P_2s'
>
> n=366
> nt=290
> neval=n-nt
> rep=100
> errlin=dim(rep)
> accuracy=dim(rep)
> precision=dim(rep)
> recall=dim(rep)
> for (k in 1:rep)
+ {
+   train=sample(1:n,nt)
+   kcweather.nb = naiveBayes(Events~.,nbdata[train,])
+   nbdata.test = nbdata[-train,1:3]
+   predict(kcweather.nb, nbdata.test, type="raw")
+   tablin=table(predict(kcweather.nb,nbdata.test,type="class"),nbdata[-train,4])
+   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])
+ }
There were 50 or more warnings (use warnings() to see the first 50)
> cat("Accuracy: ",mean(accuracy))
Accuracy:  0.6040789> cat("Precision :  ",mean(precision))
Precision :   0.6561584> cat("Recall: ",mean(recall))
Recall:  0.7884978>
```

Accuracy: 0.6040789

Precision: 0.65615

Recall: 0.7884978

Now we shall have all predictors as quantitative predictor:

```
> nbdata=data[,c("Temp.F","Humidity.percentage","Precip.in","Events")]
> n=366
> nt=290
> neval=n-nt
> rep=100
> errlin=dim(rep)
> accuracy=dim(rep)
> precision=dim(rep)
> recall=dim(rep)
> for  (k in 1:rep)
+ {
+   train=sample(1:n,nt)
+   kcweather.nb = naiveBayes(Events~.,nbdata[train,])
+   nbdata.test = nbdata[-train,1:3]
+   predict(kcweather.nb, nbdata.test, type="raw")
+   tablin=table(predict(kcweather.nb,nbdata.test,type="class"),nbdata[-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])
+ }

                    Rain Rain_Thunderstorm Snow
  Rain                21                 7    0
  Rain_Thunderstorm    8                21    0
  Snow                 7                 0   12

                    Rain Rain_Thunderstorm Snow
  Rain                25                12    0
  Rain_Thunderstorm    5                21    0
  Snow                 5                 0    8
```

```
                     Rain Rain_Thunderstorm Snow
  Rain                 28                 7    0
  Rain_Thunderstorm     6                22    0
  Snow                  1                 0   12

                     Rain Rain_Thunderstorm Snow
  Rain                 25                 6    0
  Rain_Thunderstorm     8                23    0
  Snow                  2                 0   12

                     Rain Rain_Thunderstorm Snow
  Rain                 27                 9    0
  Rain_Thunderstorm     4                24    0
  Snow                  5                 0    7

                     Rain Rain_Thunderstorm Snow
  Rain                 19                12    3
  Rain_Thunderstorm     9                16    0
  Snow                  5                 0   12

                     Rain Rain_Thunderstorm Snow
  Rain                 32                 2    0
  Rain_Thunderstorm    10                18    0
  Snow                  4                 0   10
> cat("Accuracy: ",mean(accuracy))
Accuracy:  0.6131579> cat("Precision :  ",mean(precision))
Precision :   0.7835841> cat("Recall: ",mean(recall))
Recall:  0.7666215> |
```

Accuracy: 0.6131579

Precision: 0.7835841

Recall: 0.7666215

Inference:

1. We could see that the given data set when run through Naïve Bayes model, the accuracy is good with all the predictors as quantitative and a relatively less precision when compared to few predictors.
2. If we compare Naives model to LDA,QDA and KNN, we see that QDA is a better model for the given data and Naïve is better than KNN if we have atleast one quantitative predictor.
3. If we have only temperature as the quantitative predictor then the recall value is high in Naive’s Model.