### [Question 2](https://blackboard.umkc.edu/webapps/bb-selfpeer-BB584b1cebbb41a/contentType/submissions/submission.jsp?course_id=_167602_1&content_id=_3027186_1&qNumber=1)

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.

*#import the dataset and make some changes*

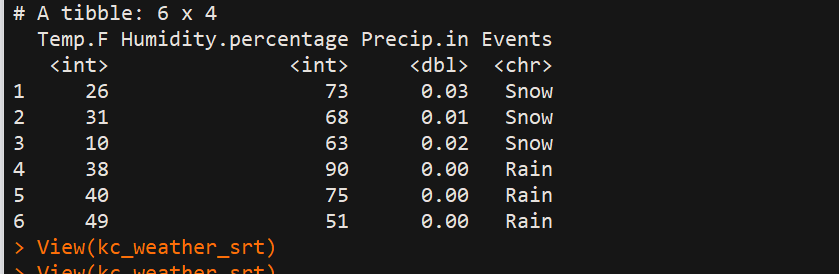
library**(**readr**)**

library**(**naivebayes**)**

kc\_weather\_srt **<-** read\_csv**(**"C:/Users/bvkka/Desktop/ISL-Deep Medhi/kc\_weather\_srt.csv"**)**

*#consider only 3 predictors - Temp.F,Humidity.%,Precip.in.*

kc\_weather\_srt**=**kc\_weather\_srt**[**,c**(**2,4,8,9**)]**



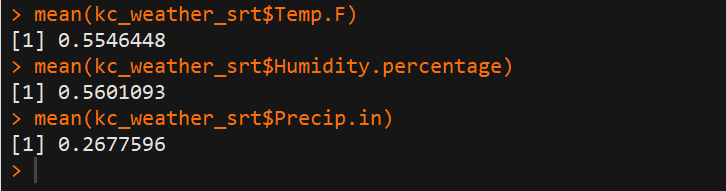
head**(**kc\_weather\_srt**)** *# to check the new data consisting of only three predictors*

***###\*\*\*\*\*categorization of three predictors\*\*\*\*###***

mean**(**kc\_weather\_srt**$**Temp.F**)** *## we see 58.74044*

mean**(**kc\_weather\_srt**$**Humidity.percentage**)***## we see 69.85246*

mean**(**kc\_weather\_srt**$**Precip.in**)***##we see 0.1728415*



*##Now, categorize the temp.F based on mean of all the values that is values >58.74 as 1 and <58.74 as 0*

*##Now, categorize the Humidity% based on mean of all the values that is values >69.85246 as 1 and <69.85246 as 0*

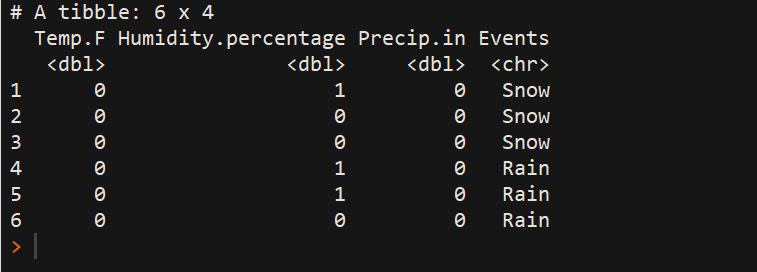
*##Now, categorize the Precip.in based on mean of all the values that is values >0.1728415 as 1 and <0.1728415 as 0*

kc\_weather\_srt**$**Temp.F**=**ifelse**(**kc\_weather\_srt**$**Temp.F**>**58.74044,1,0**)**

kc\_weather\_srt**$**Humidity.percentage**=**ifelse**(**kc\_weather\_srt**$**Humidity.percentage**>**69.85246,1,0**)**

kc\_weather\_srt**$**Precip.in**=**ifelse**(**kc\_weather\_srt**$**Precip.in**>**0.1728415,1,0**)**

head**(**kc\_weather\_srt**)**



*#replications*

rep**=**100

*# newly added*

accuracy**=**dim**(**rep**)**

precision\_snow**=**dim**(**rep**)**

precision\_rain**=**dim**(**rep**)**

precision\_rainThunderstorm**=**dim**(**rep**)**

recall\_snow**=**dim**(**rep**)**

recall\_rain**=**dim**(**rep**)**

recall\_rainThunderstorm**=**dim**(**rep**)**

*#splitting the dataset into training and test sets, also install caTools packages*

*#install.packages('caTools')*

library**(**caTools**)**

set.seed**(**123**)**

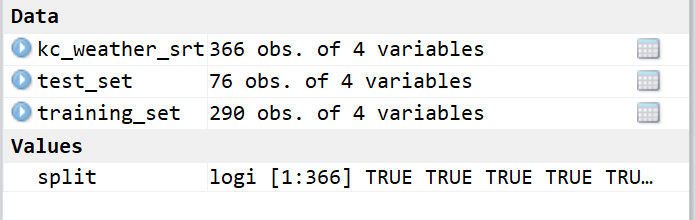
for**(**k in 1**:**rep**)**

**{**

split**=**sample.split**(**kc\_weather\_srt**$**Events,SplitRatio **=** 0.7923**)**

training\_set**=**subset**(**kc\_weather\_srt,split**==**TRUE**)**

test\_set**=**subset**(**kc\_weather\_srt,split**==**FALSE**)**



*#install.packages('e1071')*

library**(**e1071**)**

Nb**=**naiveBayes**(**formula**=**Events**~**.,data**=**training\_set**)**

summary**(**Nb**)**

*#predicting the test set results*

y\_pred**=**predict**(**Nb,newdata**=**test\_set**[-**4**])**

*#making the confusion matrix*

cm**=**table**(**y\_pred,test\_set**[**,4**])**

accuracy**[**k**]=**mean**(**y\_pred**==**test\_set**[**,4**])**

precision**=**precision**<-**diag**(**cm**)/**colSums**(**cm**)**

precision\_rainThunderstorm**[**k**]=**precision**[**3**]**

precision\_snow**[**k**]=**precision**[**2**]**

precision\_rain**[**k**]=**precision**[**1**]**

recall**=**recall**<-**diag**(**cm**/**rowSums**(**cm**))**

recall\_rainThunderstorm**[**k**]=**recall**[**3**]**

recall\_snow**[**k**]=**recall**[**2**]**

recall\_rain**[**k**]=**recall**[**1**]**

**}**

mean**(**accuracy**)** *##0.5307895*

mean**(**precision\_snow**)***##0.7293103*

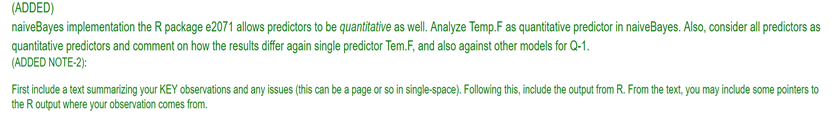
mean**(**precision\_rain**)***##0.2483784*

mean**(**precision\_rainThunderstorm**)***##1*

mean**(**recall\_snow**)***##0.7474423*

mean**(**recall\_rain**)***##0.6610176*

mean**(**recall\_rainThunderstorm**)***##0.2998356*



***#consider Temp.F as quantitative and other two predictors as qualitative***

*#import the dataset and make some changes*

library**(**readr**)**

library**(**naivebayes**)**

library**(**e2071**)**

kc\_weather\_srt **<-** read\_csv**(**"C:/Users/bvkka/Desktop/ISL-Deep Medhi/kc\_weather\_srt.csv"**)**

*#consider only 3 predictors - Temp.F,Humidity.%,Precip.in.*

kc\_weather\_srt**=**kc\_weather\_srt**[**,c**(**2,4,8,9**)]**

head**(**kc\_weather\_srt**)** *# to check the new data consisting of only three predictors*

***###\*\*\*\*\*categorization of two predictors\*\*\*\*###***

mean**(**kc\_weather\_srt**$**Humidity.percentage**)***## we see 69.85246*

mean**(**kc\_weather\_srt**$**Precip.in**)***##we see 0.1728415*

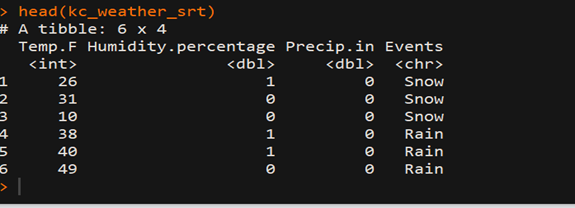
*##Now, categorize the temp.F based on mean of all the values that is values >69.85246 as 1 and <69.85246 as 0*

*##Now, also categorize the Precip on mean of all the values that is values >0.1728415 as 1 and <0.1728415 as 0*

kc\_weather\_srt**$**Humidity.percentage**=**ifelse**(**kc\_weather\_srt**$**Humidity.percentage**>**69.85246,1,0**)**

kc\_weather\_srt**$**Precip.in**=**ifelse**(**kc\_weather\_srt**$**Precip.in**>**0.1728415,1,0**)**

head**(**kc\_weather\_srt**)**



*#replications*

rep**=**100

*# newly added*

accuracy**=**dim**(**rep**)**

precision\_snow**=**dim**(**rep**)**

precision\_rain**=**dim**(**rep**)**

precision\_rainThunderstorm**=**dim**(**rep**)**

recall\_snow**=**dim**(**rep**)**

recall\_rain**=**dim**(**rep**)**

recall\_rainThunderstorm**=**dim**(**rep**)**

*#splitting the dataset into training and test sets, also install caTools packages*

*#install.packages('caTools')*

library**(**caTools**)**

set.seed**(**123**)**

for**(**k in 1**:**rep**)**

**{**

split**=**sample.split**(**kc\_weather\_srt**$**Events,SplitRatio **=** 0.7923**)**

training\_set**=**subset**(**kc\_weather\_srt,split**==**TRUE**)**

test\_set**=**subset**(**kc\_weather\_srt,split**==**FALSE**)**

*#install.packages('e1071')*

library**(**e1071**)**

Nb**=**naiveBayes**(**formula**=**Events**~**.,data**=**training\_set**)**

summary**(**Nb**)**

*#predicting the test set results*

y\_pred**=**predict**(**Nb,newdata**=**test\_set**[-**4**])**

*#making the confusion matrix*

cm**=**table**(**y\_pred,test\_set**[**,4**])**

accuracy**[**k**]=**mean**(**y\_pred**==**test\_set**[**,4**])**

precision**=**precision**<-**diag**(**cm**)/**colSums**(**cm**)**

precision\_rainThunderstorm**[**k**]=**precision**[**3**]**

precision\_snow**[**k**]=**precision**[**2**]**

precision\_rain**[**k**]=**precision**[**1**]**

recall**=**recall**<-**diag**(**cm**/**rowSums**(**cm**))**

recall\_rainThunderstorm**[**k**]=**recall**[**3**]**

recall\_snow**[**k**]=**recall**[**2**]**

recall\_rain**[**k**]=**recall**[**1**]**

**}**

mean**(**accuracy**)** *##0.6923684*

mean**(**precision\_snow**)***##0.7682759*

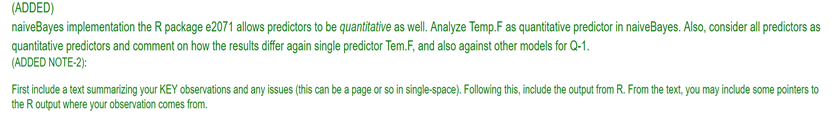
mean**(**precision\_rain**)***##0.5497292*

mean**(**precision\_rainThunderstorm**)***##1*

mean**(**recall\_snow**)***##0.7787254*

mean**(**recall\_rain**)***##0.7723869*

mean**(**recall\_rainThunderstorm**)***##0.4878522*



*#consider all predictors as quantitative and use Naive Bayes Classifier and install package (e2071)*

*#import the dataset and make some changes*

library**(**readr**)**

library**(**naivebayes**)**

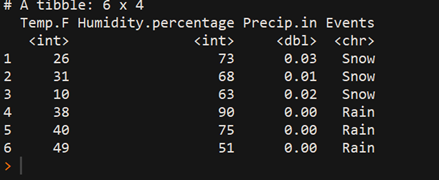
library**(**e2071**)**

kc\_weather\_srt **<-** read\_csv**(**"C:/Users/bvkka/Desktop/ISL-Deep Medhi/kc\_weather\_srt.csv"**)**

*#consider only 3 predictors - Temp.F,Humidity.%,Precip.in.*

kc\_weather\_srt**=**kc\_weather\_srt**[**,c**(**2,4,8,9**)]**

head**(**kc\_weather\_srt**)**



*#replications*

rep**=**100

*# newly added*

*#snow=1 rain=0 thunderstorm=2*

accuracy**=**dim**(**rep**)**

precision\_snow**=**dim**(**rep**)**

precision\_rain**=**dim**(**rep**)**

precision\_rainThunderstorm**=**dim**(**rep**)**

recall\_snow**=**dim**(**rep**)**

recall\_rain**=**dim**(**rep**)**

recall\_rainThunderstorm**=**dim**(**rep**)**

*#splitting the dataset into training and test sets, also install caTools packages*

*#install.packages('caTools')*

library**(**caTools**)**

set.seed**(**123**)**

for**(**k in 1**:**rep**)**

**{**

split**=**sample.split**(**kc\_weather\_srt**$**Events,SplitRatio **=** 0.7923**)**

training\_set**=**subset**(**kc\_weather\_srt,split**==**TRUE**)**

test\_set**=**subset**(**kc\_weather\_srt,split**==**FALSE**)**

*#install.packages('e1071')*

library**(**e1071**)**

Nb**=**naiveBayes**(**formula**=**Events**~**.,data**=**training\_set**)**

summary**(**Nb**)**

*#predicting the test set results*

y\_pred**=**predict**(**Nb,newdata**=**test\_set**[-**4**])**

*#making the confusion matrix*

cm**=**table**(**y\_pred,test\_set**[**,4**])**

accuracy**[**k**]=**mean**(**y\_pred**==**test\_set**[**,4**])**

precision**=**precision**<-**diag**(**cm**)/**colSums**(**cm**)**

precision\_rainThunderstorm**[**k**]=**precision**[**3**]**

precision\_snow**[**k**]=**precision**[**2**]**

precision\_rain**[**k**]=**precision**[**1**]**

recall**=**recall**<-**diag**(**cm**/**rowSums**(**cm**))**

recall\_rainThunderstorm**[**k**]=**recall**[**3**]**

recall\_snow**[**k**]=**recall**[**2**]**

recall\_rain**[**k**]=**recall**[**1**]**

**}**

mean**(**accuracy**)** *##0.7507895*

mean**(**precision\_snow**)***##0.73620689*

mean**(**precision\_rain**)***##0.7064865*

mean**(**precision\_rainThunderstorm**)***##0.957*

mean**(**recall\_snow**)***##0.7498165*

mean**(**recall\_rain**)***##0.766421*

mean**(**recall\_rainThunderstorm**)***##0.7367071*

**Summary (Naïve Bayes):**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Naive Bayes | Accuracy | Precision Snow | Precision Rain | Precision Rain Thunderstorm | Recall Snow | Recall Rain | Recall Thunderstorm |
| Three Qualitative Predictors | 0.5307895 | 0.7293103 | 0.2483784 | 1 | 0.7474423 | 0.6610176 | 0.2998356 |
| Only temperature Quantitative | 0.6923684 | 0.7682759 | 0.5497292 | 1 | 0.7787254 | 0.7723869 | 0.4878522 |
| All Three  Quantitative predictors | 0.7507895 | 0.73620689 | 0.7064865 | 0.957 | 0.7498165 | 0.766421 | 0.7367071 |

**Text Summarization:**

1. When all the three predictors i.e. Temperature, Humidity and Precipitation are categorized as Qualitative, we observe that the Accuracy using Naïve Bayes Model is very less. The Recall values are also less but the Precision values are good, and we see the Precision of Rain happening is 100%.
2. When only the temperature is quantitative, and rest two predictors are quantitative, the recall value is high compared to the remaining models and has better accuracy and a good precision values. So, when recall and accuracy are of higher importance then this model is the best fit.
3. When all three Predictors are Quantitative, we observe that the accuracy is increased (is best) and the Precision and recall values are also good. So, from above three summary, we can conclude that the predictors being quantitative is more better when we use Naïve Bayes compared to when we use this model for Qualitative Predictors.
4. When Naïve Bayes model results is compared to results obtained for LDA, QDA and KNN model the LDA, QDA model dominates the Naïve Bayes model in all the metrics accuracy, precision and recall values.
5. The Naïve Bayes model performs better than the KNN model (K=5) when at least one of the predictor is Quantitative.