ABSTRACT

The skin segmentation dataset is constructed over blue, green, red color space. The skin dataset is collected by randomly sampling B, G, R values from face images of various age groups (young, middle, and old), race groups (white, black, and Asian), and genders obtained from FERET database and PAL database. Total learning sample size is 245057; out of which 50859 is skin samples and 194198 is non-skin samples.

OBJECTIVE

Prediction of skin and non-skin can be done based on the evaluation of blue, green and red pixels.

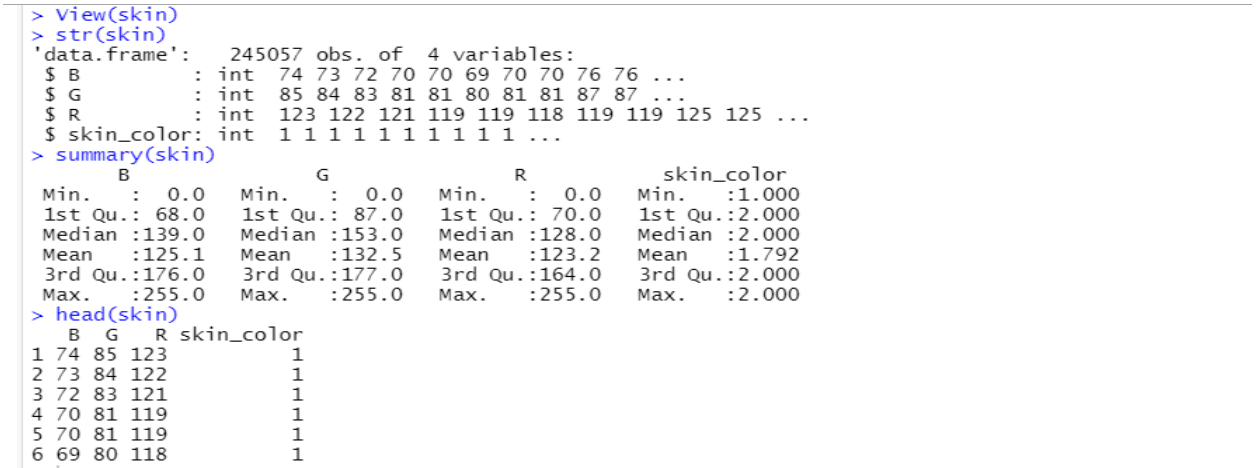
LOADING DATASET

```
skin<- read.csv("c:/data/Skin_Nonskin.csv")
```

DATA EXPLORATION

Explore data to get an idea about its structure

```
View(skin)
str(skin)              #displays internal structure
summary(skin)          #displays descriptive statistics of every variable in the dataset
head(skin)             #displays top 6 values
```

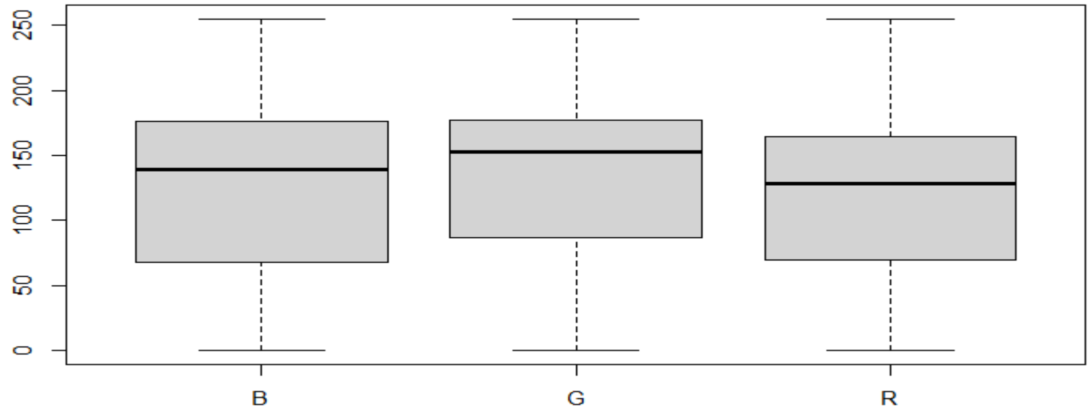


```
> View(skin)
> str(skin)
'data.frame':   245057 obs. of  4 variables:
 $ B         : int  74 73 72 70 70 69 70 70 76 76 ...
 $ G         : int  85 84 83 81 81 80 81 81 87 87 ...
 $ R         : int  123 122 121 119 119 118 119 119 125 125 ...
 $ skin_color: int  1 1 1 1 1 1 1 1 1 1 ...
> summary(skin)
       B               G               R           skin_color
 Min.   :  0.0   Min.   :  0.0   Min.   :  0.0   Min.   :1.000
 1st Qu.: 68.0   1st Qu.: 87.0   1st Qu.: 70.0   1st Qu.:2.000
 Median :139.0   Median :153.0   Median :128.0   Median :2.000
 Mean   :125.1   Mean   :132.5   Mean   :123.2   Mean   :1.792
 3rd Qu.:176.0   3rd Qu.:177.0   3rd Qu.:164.0   3rd Qu.:2.000
 Max.   :255.0   Max.   :255.0   Max.   :255.0   Max.   :2.000
> head(skin)
   B  G   R skin_color
1 74 85 123          1
2 73 84 122          1
3 72 83 121          1
4 70 81 119          1
5 70 81 119          1
6 69 80 118          1
```

DATA VISUALIZATION

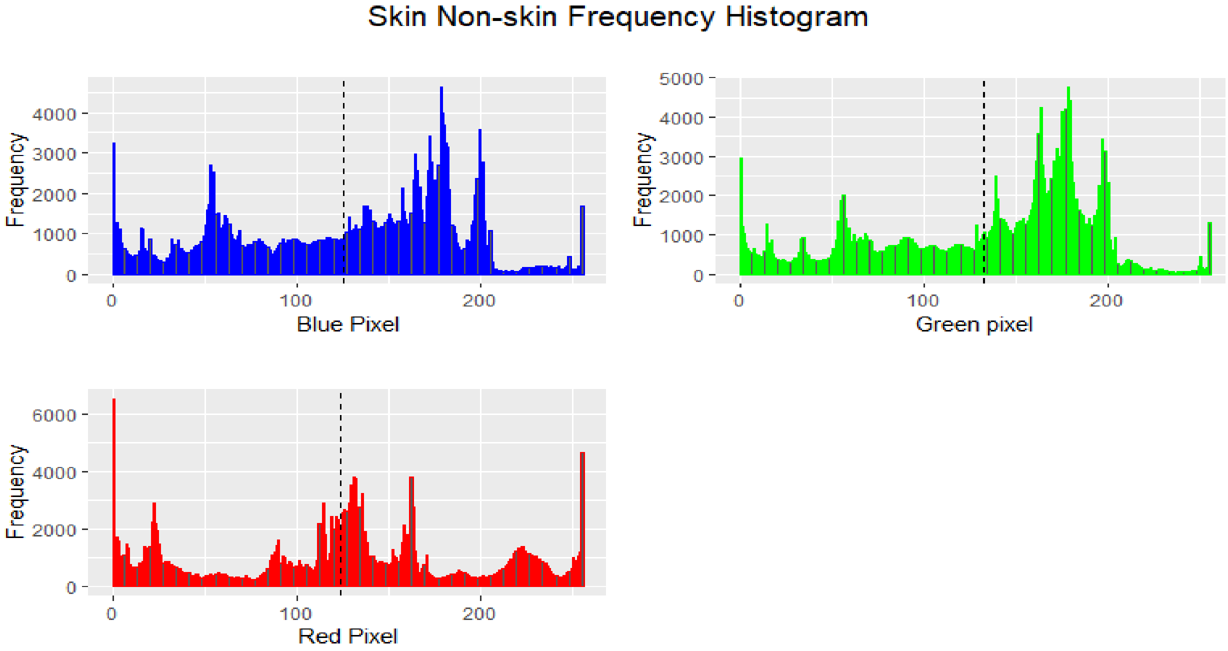
Box plot representation is shown below:

```
boxplot(skin[,1:3])
```



It is used to represent descriptive statistics of each variable in a dataset. It represents the minimum, first quartile, median, third quartile, and the maximum values of a variable.

Histogram representation of data:



The blue, green , red components are relatively normally distributed.

DATA ANALYTICS TECHNIQUES

1. NAIVE BAYES

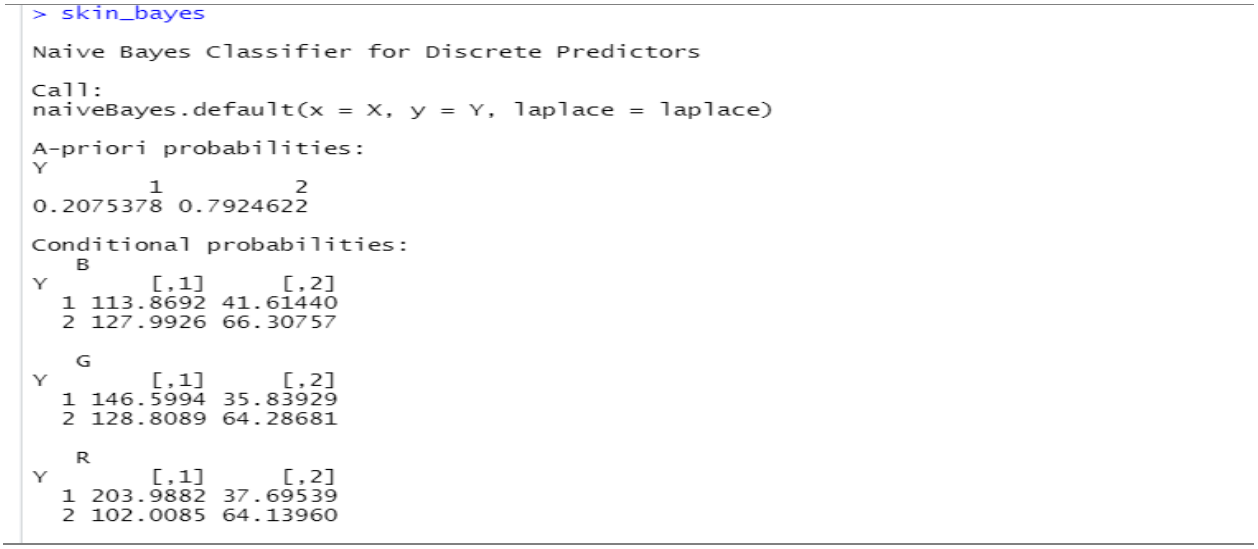
Here the dataset contains 245057 samples. It has different blue, green, red pixels to identify whether it is skin or non-skin. For classifying this Naïve Bayes classification technique can be used.

First step is to partition the dataset as training and testing data. 'caTools 'package is used to make a balanced partitioning.

```
sample_data = sample.split(skin, SplitRatio = 0.75)    #splitting the rows
train <- subset(skin, sample_data == TRUE)             #logic TRUE values move to train data
test <- subset(skin, sample_data == FALSE)             #logic FALSE data move to test
```

Now utilize 'e1071' package for classifying

```
set.seed(120)                                          # set seed function get same result each time
skin_bayes <- naiveBayes(skin_color ~ ., data = train)
skin_bayes
```



```
> skin_bayes

Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
        1         2 
0.2075378 0.7924622 

Conditional probabilities:
   B
Y       [,1]     [,2]
  1 113.8692 41.61440
  2 127.9926 66.30757

   G
Y       [,1]     [,2]
  1 146.5994 35.83929
  2 128.8089 64.28681

   R
Y       [,1]     [,2]
  1 203.9882 37.69539
  2 102.0085 64.13960
```

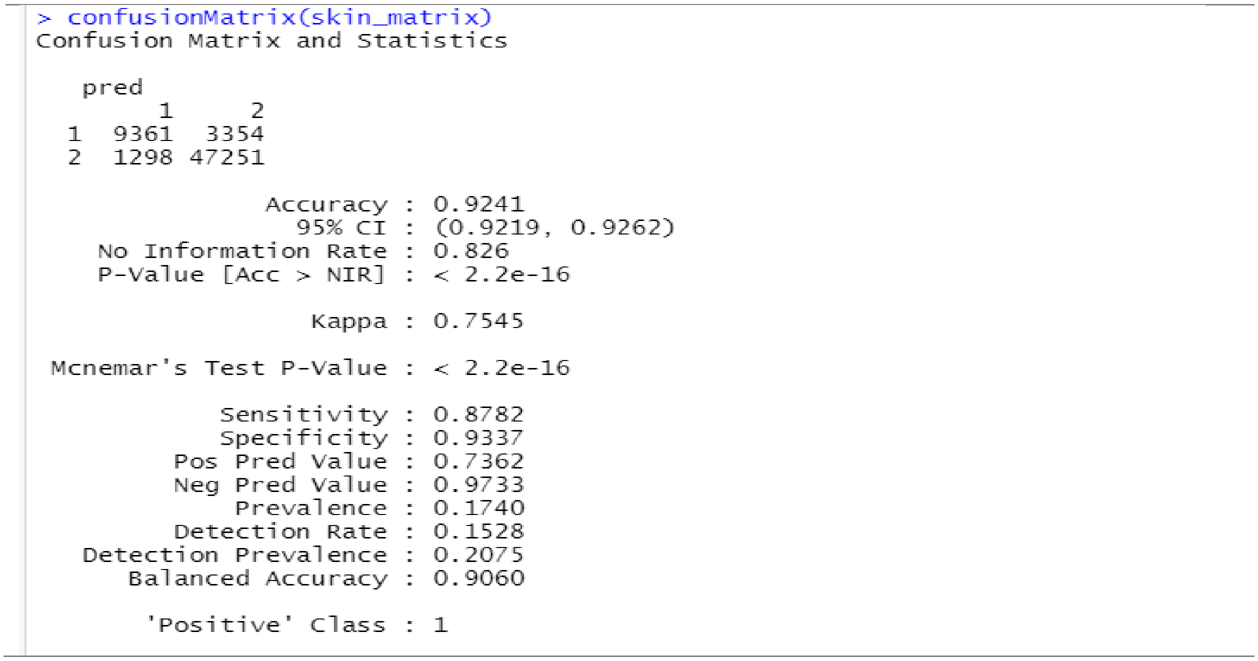
The Y values are the means and standard deviations of the predictors within each class.

Prediction is done on test data because higher test accuracy is found on test data.

```
pred <- predict(skin_bayes, newdata = test)
```

Accuracy is determined using confusion matrix. 'caret' package is used for confusion matrix.

```
skin_matrix <- table(test$skin_color, pred)
confusionMatrix(skin_matrix)                          #Model evaluation
```



```
> confusionMatrix(skin_matrix)
Confusion Matrix and Statistics

   pred
        1     2
  1  9361  3354
  2  1298 47251

               Accuracy : 0.9241
                 95% CI : (0.9219, 0.9262)
    No Information Rate : 0.826
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7545

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8782
            Specificity : 0.9337
         Pos Pred Value : 0.7362
         Neg Pred Value : 0.9733
             Prevalence : 0.1740
         Detection Rate : 0.1528
   Detection Prevalence : 0.2075
      Balanced Accuracy : 0.9060

       'Positive' Class : 1
```

RESULT

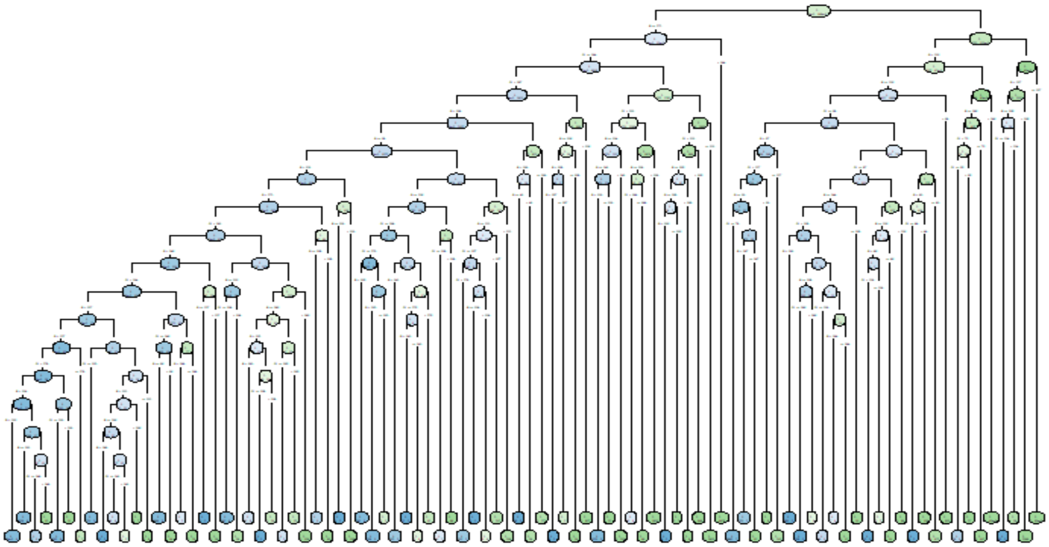
In the above result, from the prediction table, 9361 out of 10659 are able to be classified as skin and 47251 out of 50605 are able to be classified as non-skin. So, the accuracy to predict skin is 87.8% and the accuracy to predict non skin is 93.37%. This results in an overall accuracy of 92.41%.

2. DECISION TREE

Another classification method Decision tree is used for classifying skin and non-skin.

For implementing decision tree ‘rpart’ package is imported and ‘rpart.plot’ package will help to get a visual plot of a decision tree.

```
skin_decision <- rpart(formula = skin_color ~.,
                     data = train,
                     method = "class",
                     control = rpart.control(cp = 0),
                     parms = list(split = "information"))
rpart.plot(skin_decision,type= 4 , extra=1)
```

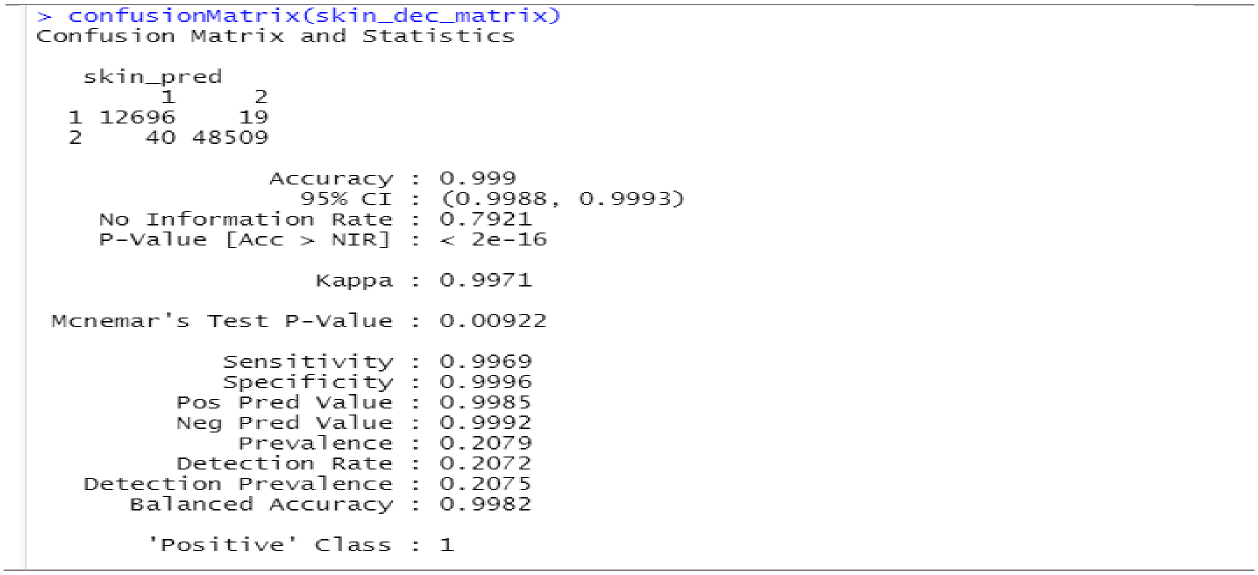


Predict the dataset as skin or non-skin using predict() on test.

```
skin_pred <- predict(object = skin_decision,
                     newdata = test,
                     type = "class")
```

Here also the accuracy is determined using confusion matrix.

```
skin_dec_matrix <- table(test$skin_color, skin_pred)
confusionMatrix(skin_dec_matrix)                          #model evaluation
```



```
> confusionMatrix(skin_dec_matrix)
Confusion Matrix and Statistics

   skin_pred
         1     2
  1 12696    19
  2    40 48509

               Accuracy : 0.999
                 95% CI : (0.9988, 0.9993)
    No Information Rate : 0.7921
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.9971

 Mcnemar's Test P-Value : 0.00922

            Sensitivity : 0.9969
            Specificity : 0.9996
         Pos Pred Value : 0.9985
         Neg Pred Value : 0.9992
             Prevalence : 0.2079
         Detection Rate : 0.2072
   Detection Prevalence : 0.2075
      Balanced Accuracy : 0.9982

       'Positive' Class : 1
```

RESULT

In the above result, from the prediction table, 12696 out of 12736 are able to be classified as skin and 48509 out of 48528 are able to be classified as non-skin. So, the accuracy to predict skin is 99.69% and the accuracy to predict non skin is 99.96%. This results in an overall accuracy of 99.90%.

CONCLUSION

The accuracy of Naïve Bayes to predict skin and non-skin on bases of green, blue, red is 92.41% and for Decision tree is 99.90%. From the overall result it is clear that Decision tree is more accurate compared to Naïve Bayes.