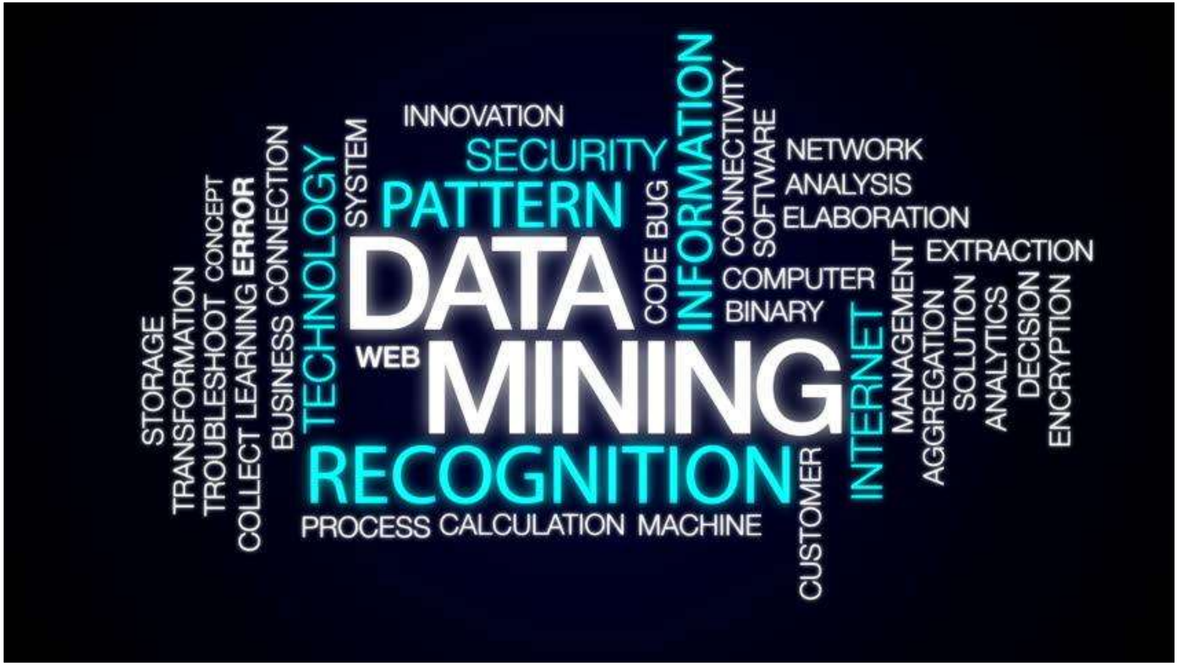


Submitted By:

Pragya Mishra

Mansi Agarwal

Sahil Bhange

Table of Contents

1. Background

A person makes a doctor’s appointment, receives all the instructions and possible assistance from doctor’s office, agrees on a time for appointment, but on the day of appointment - they doesn’t show up. Our dataset contains real world medical appointments records of the capital city of Espirito Santo State - Vitoria in Brazil. In Brazil, government provides free scholarship to eligible citizens which allows them to take free medical services. Over a period of time, it has been observed that close to 30% of patients do not show up on their scheduled day of appointment. With this high rate of no-show, government & health centers are concerned & want to address this issue.

Our report focuses on analyzing various attributes impacting the patient’s ability to attend the scheduled appointment & through our data mining techniques, we are trying to predict if the health centers, can at the time of booking an appointment, judge the chances of patient’s likelihood to attend or miss the appointment. We will be using various data mining techniques like Naïve Bayes, Random Forest & Association rule mining to come up with some concrete information which can be useful for both government as well as health center.

2. Objective

Data Mining Objective -

Predicting if a particular patient will show up for a scheduled appointment or will miss it

Business Objective -

- Used by doctors, hospitals and government to understand the factors impacting patient’s ability to attend or miss an appointment
- Doctors can increase or decrease their daily appointments based the prediction of patients chances of showing up/not
- Government can determine the impact of scholarship if they need to extend or modify this benefit to citizens
- Health centers can determine if they should invest on SMS service or not
- Health centers can determine if the awaiting time affects patients’ ability to attend /miss the appointment. Accordingly, they can increase the number of doctors
- Stakeholders – Government, Doctors, Patients

3. Dataset Information

✓ **Source** – 2015-16 Public healthcare appointment data from https://www.kaggle.com/joniarroba/noshowappointments

✓ **Type** – Categorical attributes

✓ **Size** –110528 rows

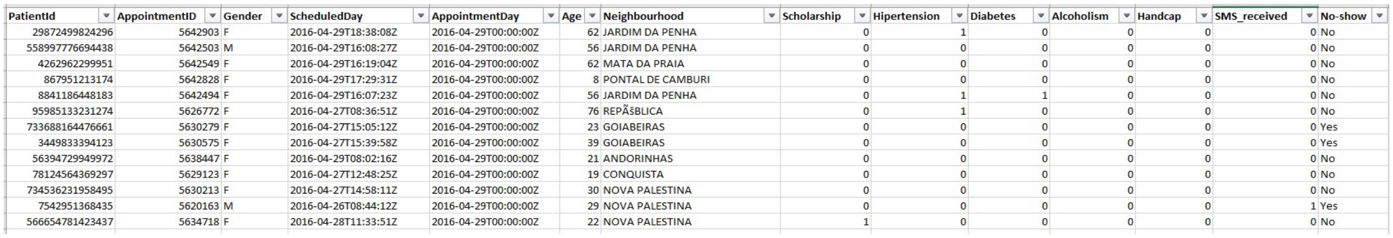
✓ **Prediction Class label –** No-show: No or Yes

✓ **Data parameters –**

- There are patients coming from 81 neighbourhood.
- The median age of patients is 37 years old.
- The median awaiting time (i.e. the time between scheduled day & appointment day) is 64 years.

✓ **Original Attributes –** 14

Snapshot of Data Set:



PatientId	AppointmentID	Gender	ScheduledDay	AppointmentDay	Age	Neighbourhood	Scholarship	Hipertension	Diabetes	Alcoholism	Handcap	SMS_received	No-show
29872499824296	5642903	F	2016-04-29T18:38:08Z	2016-04-29T00:00:00Z	62	JARDIM DA PENHA	0	1	0	0	0	0	No
558997776694438	5642503	M	2016-04-29T16:08:27Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	0	0	0	0	0	No
4262962299951	5642549	F	2016-04-29T16:19:04Z	2016-04-29T00:00:00Z	62	MATA DA PRAIA	0	0	0	0	0	0	No
867951213174	5642828	F	2016-04-29T17:29:31Z	2016-04-29T00:00:00Z	8	PONTAL DE CAMBURI	0	0	0	0	0	0	No
8841186448183	5642494	F	2016-04-29T16:07:23Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	1	1	0	0	0	No
95985133231274	5626772	F	2016-04-27T08:36:51Z	2016-04-29T00:00:00Z	76	REPÃšBLICA	0	1	0	0	0	0	No
733688164476661	5630279	F	2016-04-27T15:05:12Z	2016-04-29T00:00:00Z	23	GOIABEIRAS	0	0	0	0	0	0	Yes
3449833394123	5630575	F	2016-04-27T15:39:58Z	2016-04-29T00:00:00Z	39	GOIABEIRAS	0	0	0	0	0	0	Yes
56394729949972	5638447	F	2016-04-29T08:02:16Z	2016-04-29T00:00:00Z	21	ANDORINHAS	0	0	0	0	0	0	No
78124564369297	5629123	F	2016-04-27T12:48:25Z	2016-04-29T00:00:00Z	19	CONQUISTA	0	0	0	0	0	0	No
734536231958495	5630213	F	2016-04-27T14:58:11Z	2016-04-29T00:00:00Z	30	NOVA PALESTINA	0	0	0	0	0	0	No
7542951368435	5620163	M	2016-04-26T08:44:12Z	2016-04-29T00:00:00Z	29	NOVA PALESTINA	0	0	0	0	0	1	Yes
566654781423437	5634718	F	2016-04-28T11:33:51Z	2016-04-29T00:00:00Z	22	NOVA PALESTINA	1	0	0	0	0	0	No

4. Data Pre-Processing

Following steps were considered for preprocessing:

- ‘ScheduledDay’ is the day for which an appointment is scheduled. Currently this attribute is in a timestamp format. We created new attribute called ‘Scheduled_date’ to record all dates in MMDDYYYY format. Also, created new attributes like ‘Sch_time’ to record time of the appointment, ‘Sch_Day’ to record day of appointment like Sunday, monday etc.

- Observations in ‘Sch_time’ are converted into broad categories like Early morning, Morning, afternoon & evening.

- ‘AppointmentDay’ is the day when an appointment is taken. This attribute doesn’t need formatting when it is loaded in R, thu we have just renamed it to ‘Appointment_date’. We have created new attribute called ‘Appt_Day’ to record the day when an appointment is taken.

- Observations in ‘‘Appt_Day’ are converted into weekdays - Sunday, Monday etc.

- Another attribute called ‘Awaiting time’ is calculated as the difference between ‘Scheduled_date’ & ‘Appointment_date’. We found that for some observations this calculated attribute was coming as negative, which doesn’t make sense, as the day when an appointment is taken cannot be later the day for which appointment is scheduled. Thus, such observations have been deleted.

- Observations in ‘‘Awaiting time’ are converted into broad categories like Same day, 1 week, 1 month, three months and six months.

- For ‘Age’ attribute, we found that some observations have negative age, which have been removed.

- Observations in ‘Age’ are converted into broad categories like Infant, Child, young, middle_aged, senior & old. Also, some observations had age as zero, which we have considered as infant.

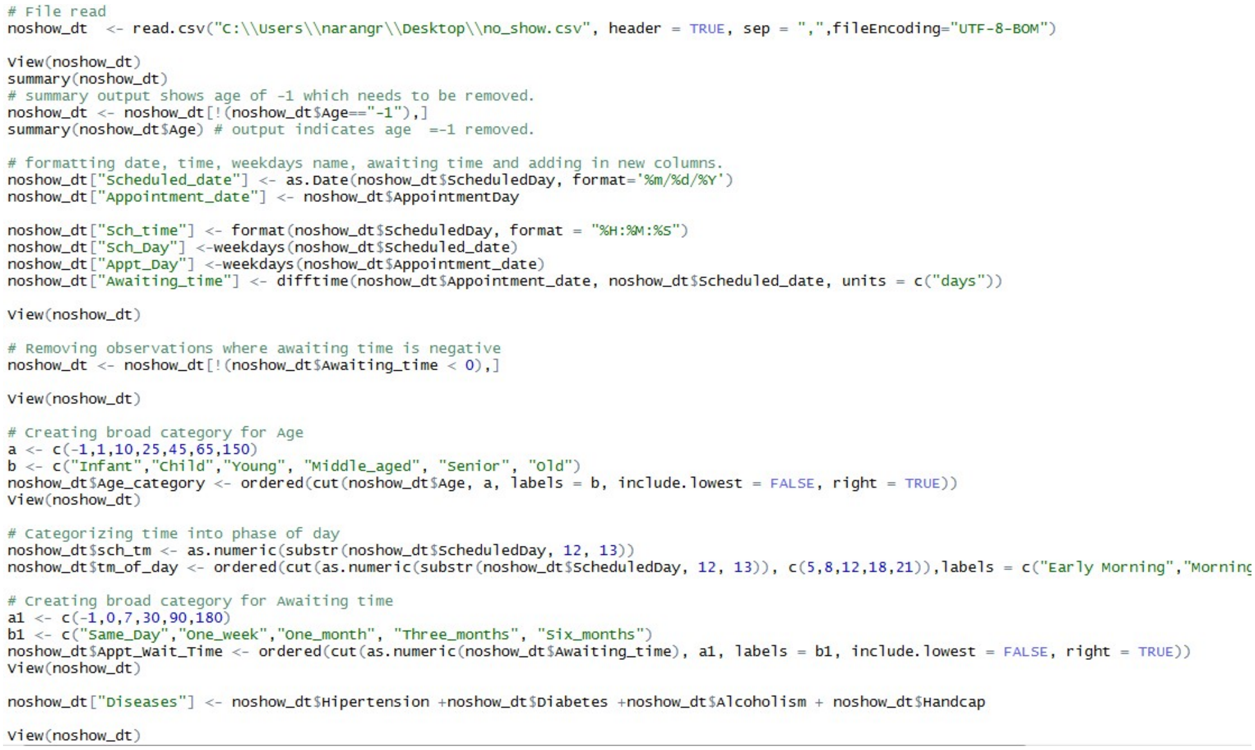
- There were no missing data in our dataset.

- We created attribute called ‘Sum of diseases’ to record the total sum of all diseases one patient has.

- For single patient having multiple appointments, we have created another attribute to record their probability of show/no show. If a patient has attended 6 appointments in past out of 10, then his chances of attending are higher than missing. For such patients, all other records have been removed & their label has been taken as ‘Show’.

- We have created factor variables for Gender, Age category, Awaiting time, sum of all diseases and Scholarship to make them binary attributes.

Preprocessing done in R

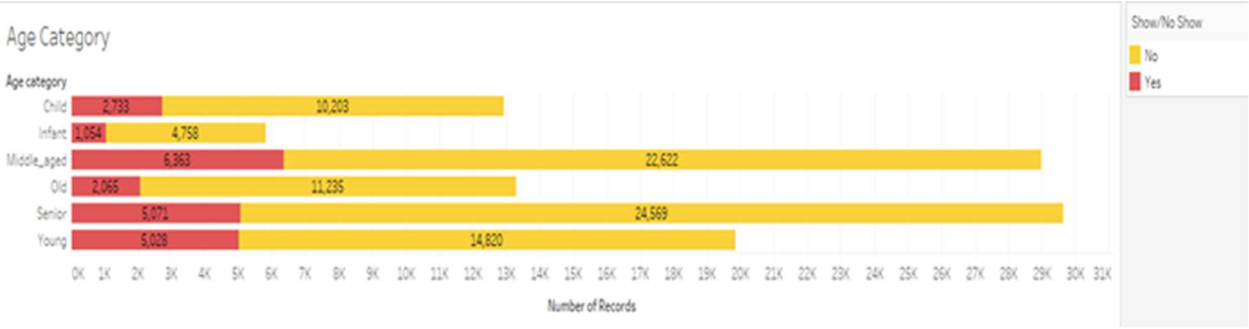


```
# File read
noshow_dt  <- read.csv("C:\\Users\\narangr\\Desktop\\no_show.csv", header = TRUE, sep = ",",fileEncoding="UTF-8-BOM")

View(noshow_dt)
summary(noshow_dt)
# summary output shows age of -1 which needs to be removed.
noshow_dt <- noshow_dt[!(noshow_dt$Age=="-1"),]
summary(noshow_dt$Age) # output indicates age  =-1 removed.

# formatting date, time, weekdays name, awaiting time and adding in new columns.
noshow_dt["Scheduled_date"] <- as.Date(noshow_dt$ScheduledDay, format='%m/%d/%Y')
noshow_dt["Appointment_date"] <- noshow_dt$AppointmentDay

noshow_dt["Sch_time"] <- format(noshow_dt$ScheduledDay, format = "%H:%M:%S")
noshow_dt["Sch_Day"] <-weekdays(noshow_dt$Scheduled_date)
noshow_dt["Appt_Day"] <-weekdays(noshow_dt$Appointment_date)
noshow_dt["Awaiting_time"] <- difftime(noshow_dt$Appointment_date, noshow_dt$Scheduled_date, units = c("days"))

View(noshow_dt)

# Removing observations where awaiting time is negative
noshow_dt <- noshow_dt[!(noshow_dt$Awaiting_time < 0),]

View(noshow_dt)

# Creating broad category for Age
a <- c(-1,1,10,25,45,65,150)
b <- c("Infant","Child","Young", "Middle_aged", "Senior", "Old")
noshow_dt$Age_category <- ordered(cut(noshow_dt$Age, a, labels = b, include.lowest = FALSE, right = TRUE))
View(noshow_dt)

# Categorizing time into phase of day
noshow_dt$sch_tm <- as.numeric(substr(noshow_dt$ScheduledDay, 12, 13))
noshow_dt$tm_of_day <- ordered(cut(as.numeric(substr(noshow_dt$ScheduledDay, 12, 13)), c(5,8,12,18,21)),labels = c("Early Morning","Morning

# Creating broad category for Awaiting time
a1 <- c(-1,0,7,30,90,180)
b1 <- c("Same_Day","One_week","One_month", "Three_months", "Six_months")
noshow_dt$Appt_Wait_Time <- ordered(cut(as.numeric(noshow_dt$Awaiting_time), a1, labels = b1, include.lowest = FALSE, right = TRUE))
View(noshow_dt)

noshow_dt["Diseases"] <- noshow_dt$Hipertension +noshow_dt$Diabetes +noshow_dt$Alcoholism + noshow_dt$Handcap

View(noshow_dt)
```

5. Data Visualization

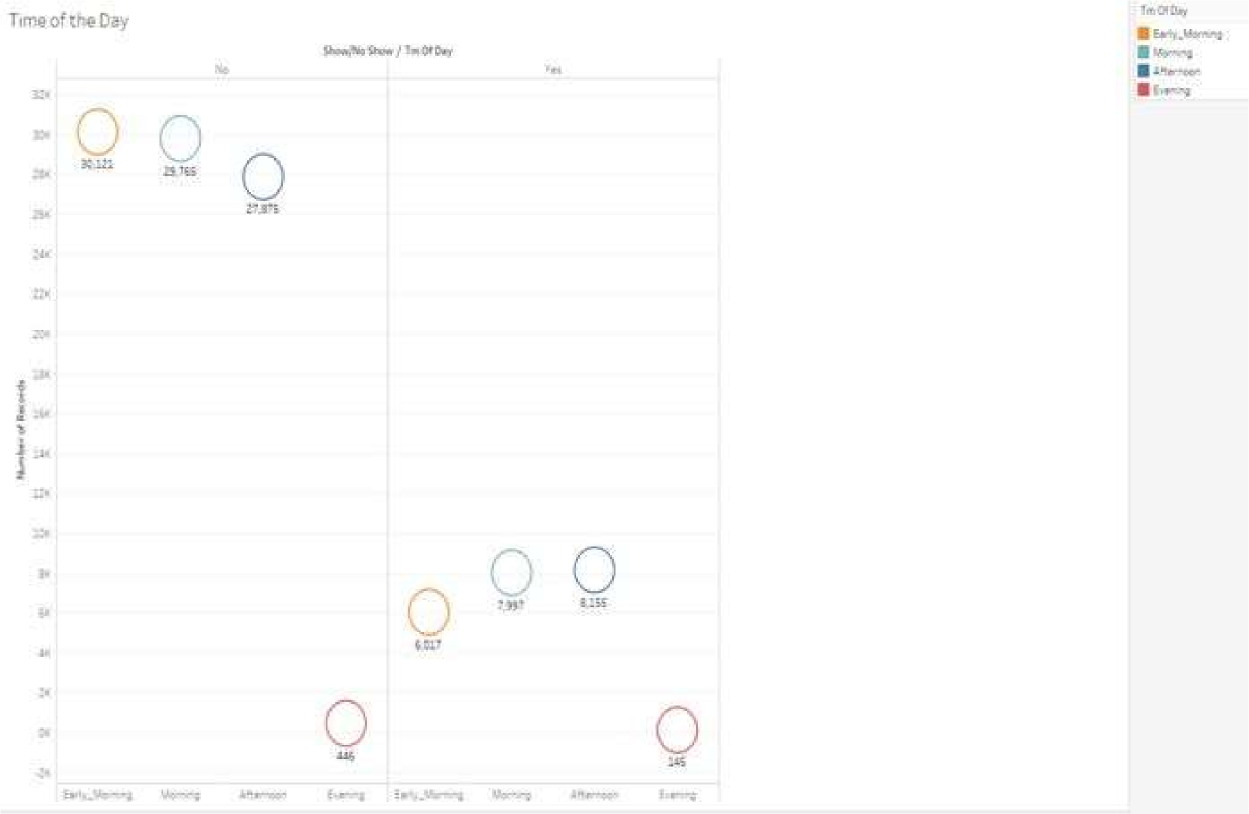
We visualized attributes on various parameters to see prominent relationships of each attribute with our class label – show/no show.

Chart1. Age Category vs Show/No_Show



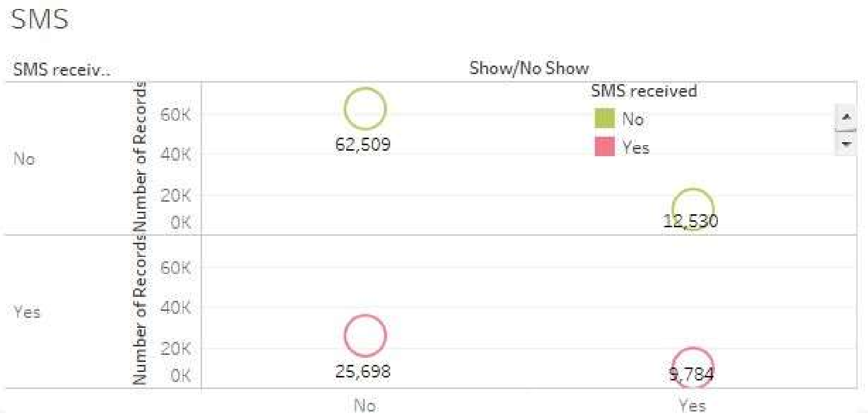
This graph tells about the number of people with show/no_show indicator based on age groups. It looks that Senior and Old aged category people are more consistent in showing up

Chart2. Time of the Day vs Show/No_Show



We see that maximum people show up in the early morning appointments and it's likely that less number of people would not show up in the early morning appointments as compared to the rest of day.

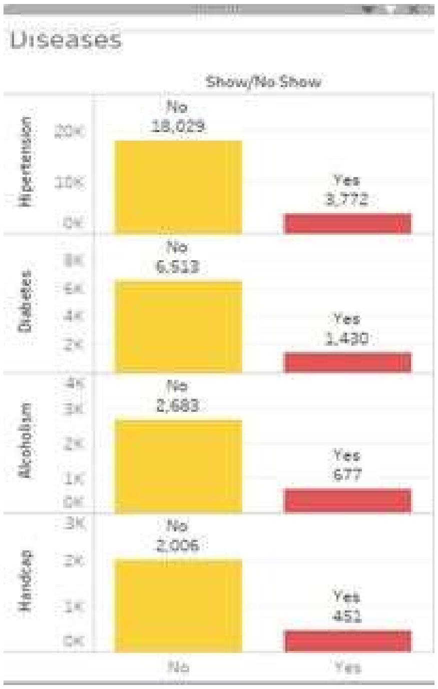
Chart3. SMS received vs Show/No_Show



This graph does a comparison on number of people showing up or not showing up depending on if they received a SMS alert about the appointment. There were around ~75k who did not receive any SMS and

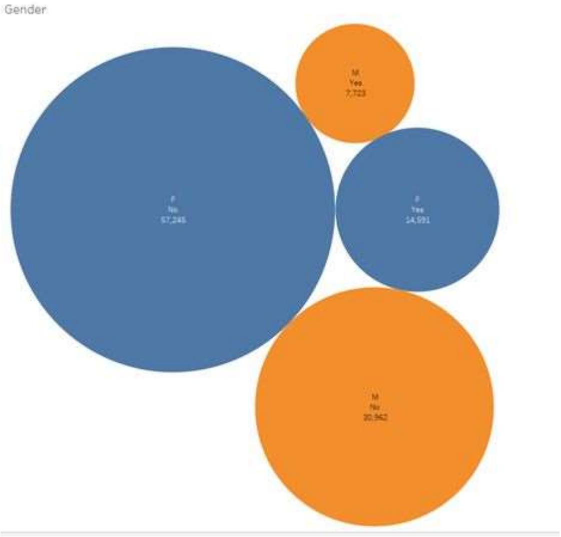
~63K (83%) showed up where as there were around ~35.5k which got the SMS and out of which ~9.7k (73%) showed up. So, it didn’t make much of difference by sending an alert through SMS

Chart4. Diseases Vs Show/No_show



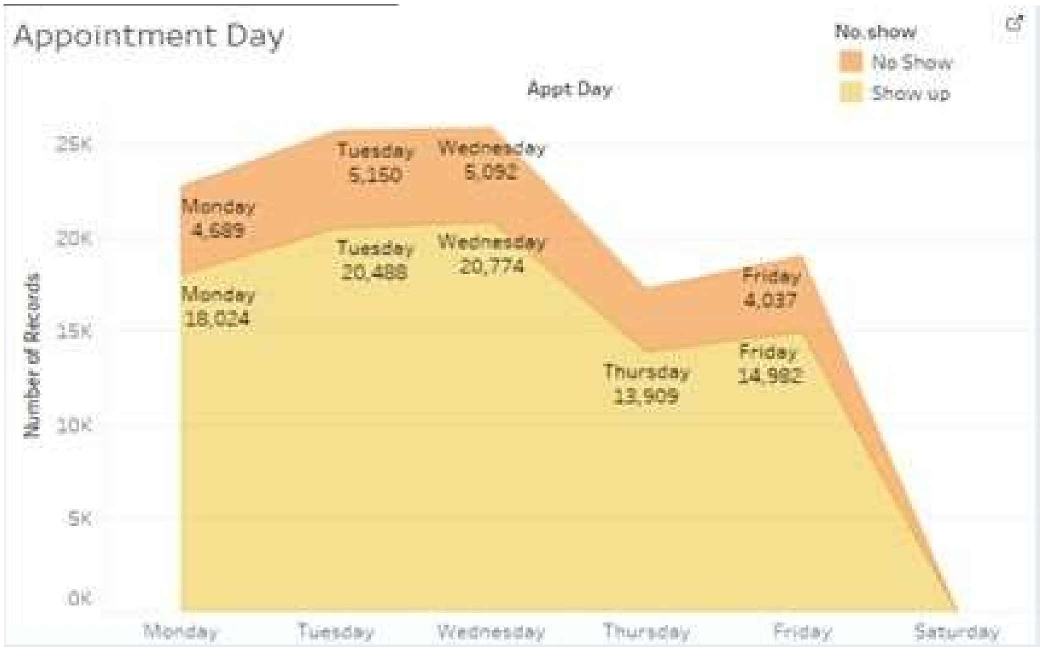
The above graph shows a record of people showing up and not showing up based on 4 different reasons of appointments like Diabetes, Alcoholism, Hypertension and Handicap.

Chart5. Gender vs Show/No_Show



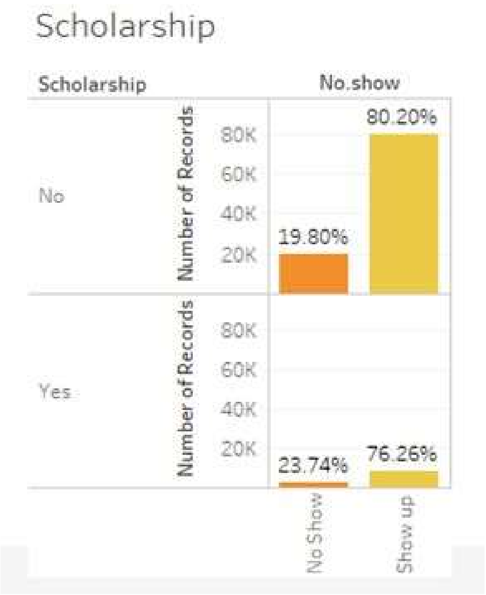
This indicates that majority of the appointments were made by females but around 70% of females show up in appointments in comparison to 80% of male population showing up

Chart6. Appointment Day Vs Show/No_Show



It looks like probability of people not showing up is consistent from Monday till Wednesday and it tends to change in the second half of week and major of them are scheduled for Tuesdays and Wednesdays

Chart7. Scholarship vs Show/No_Show



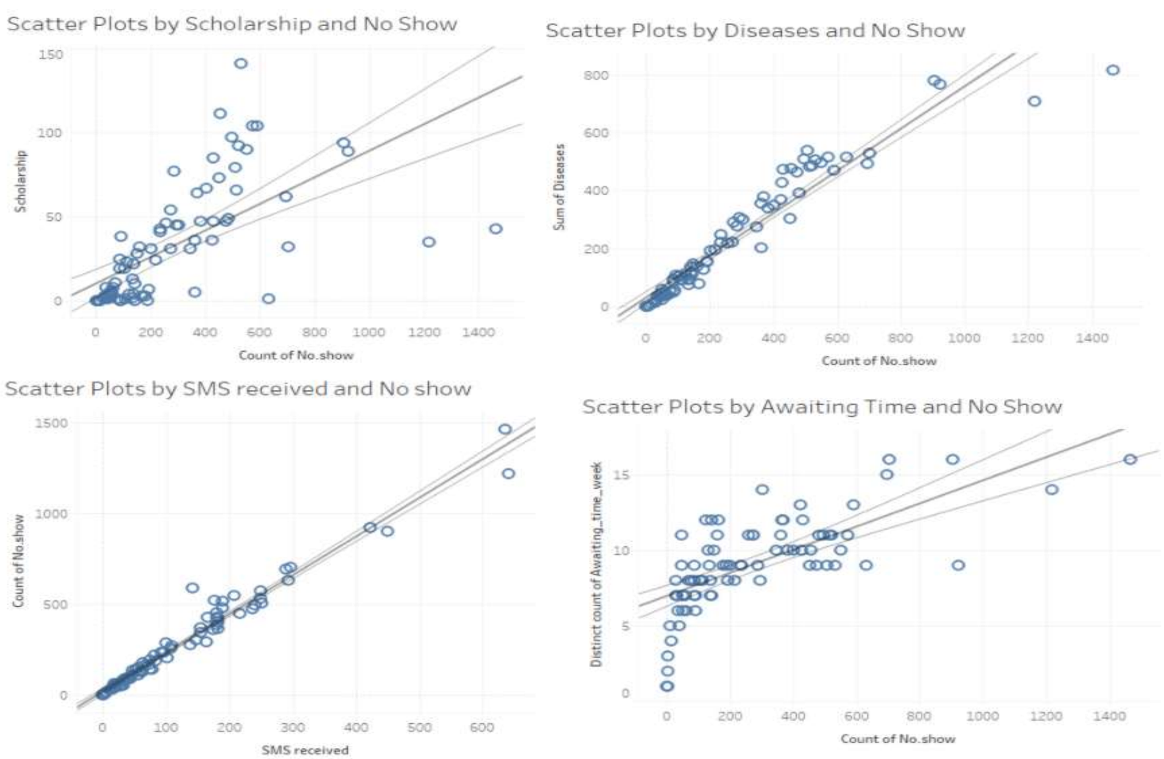
This indicates that it’s really not making any difference in not showing up to the appointments made even people had Scholarships.

<u>Chart8. Pearson Correlation charts</u>

Pearson's correlation coefficient (r) is a measure of the strength of the association between any two variables. The first step in studying the relationship between two continuous variables is to draw a scatter plot of the variables to check for linearity .

A scatter plot is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data. If the points are color-coded, one additional variable can be displayed. The data are displayed as a collection of points, each having the value of one variable determining the position on the horizontal axis and the value of the other variable determining the position on the vertical axis

We have taken scatter plots to show distribution of people not showing up depending on different attributes like Scholarship, Diseases, SMS received and Awaiting Time and their pearson correlation to to measure the strength of correlation. Pearson's correlation coefficient (r) for continuous (interval level) data ranges from -1 to +1



We can see here that there is a positive correlation of NO Show with all the other attributes we have considered.

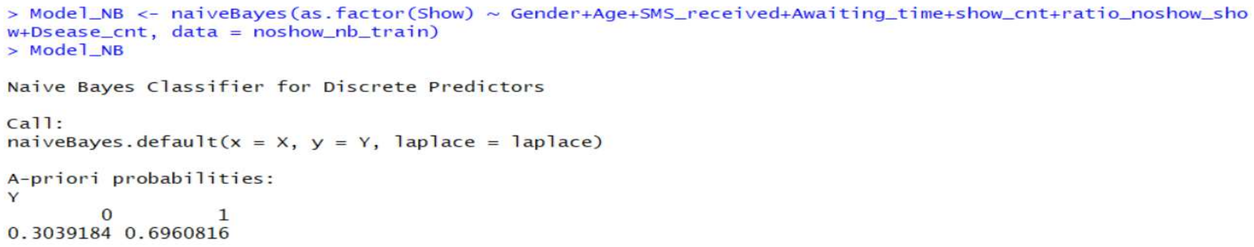
The Sum of Disease is an important factor that decides the number of no show with correlation factor very close to 1 (0.9).

6. Model Application

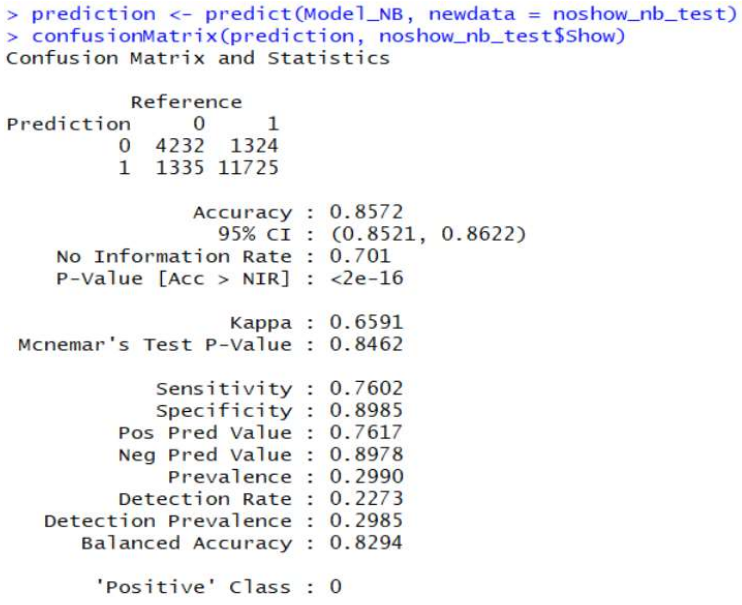
Model 1 -Naïve Bayes

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. For example, a fruit may be considered to be an apple if it is red, round, and about 10 cm in diameter. A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of any possible correlations between the color, roundness, and diameter features.

Naïve Bayes Code in R

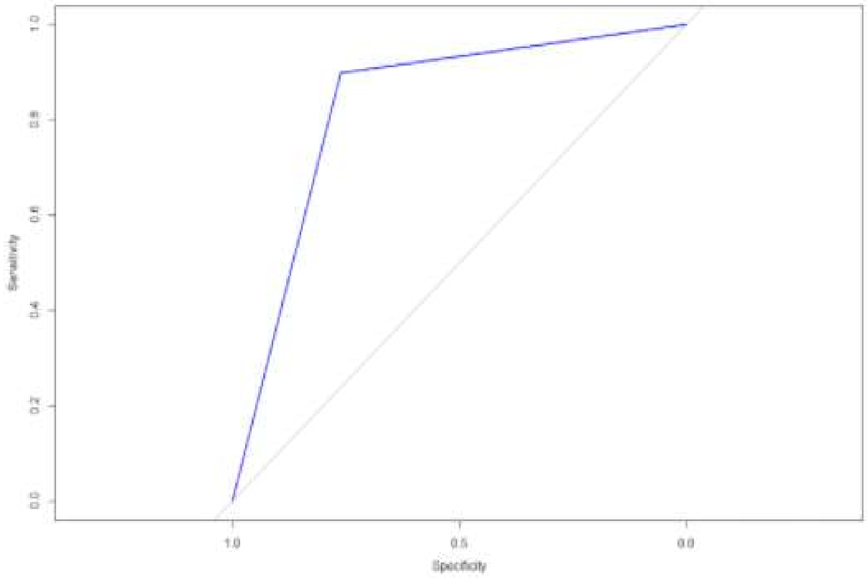


```
> Model_NB <- naiveBayes(as.factor(Show) ~ Gender+Age+SMS_received+Awaiting_time+show_cnt+ratio_noshow_sho
w+Dsease_cnt, data = noshow_nb_train)
> Model_NB

Naive Bayes Classifier for Discrete Predictors

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

A-priori probabilities:
Y
        0         1 
0.3039184 0.6960816 
```



```
> prediction <- predict(Model_NB, newdata = noshow_nb_test)
> confusionMatrix(prediction, noshow_nb_test$Show)
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0  4232  1324
         1  1335 11725
                                          
               Accuracy : 0.8572          
                 95% CI : (0.8521, 0.8622)
    No Information Rate : 0.701           
    P-Value [Acc > NIR] : <2e-16          
                                          
                  Kappa : 0.6591          
 Mcnemar's Test P-Value : 0.8462          
                                          
            Sensitivity : 0.7602          
            Specificity : 0.8985          
         Pos Pred Value : 0.7617          
         Neg Pred Value : 0.8978          
             Prevalence : 0.2990          
         Detection Rate : 0.2273          
   Detection Prevalence : 0.2985          
      Balanced Accuracy : 0.8294          
                                          
       'Positive' Class : 0               
```

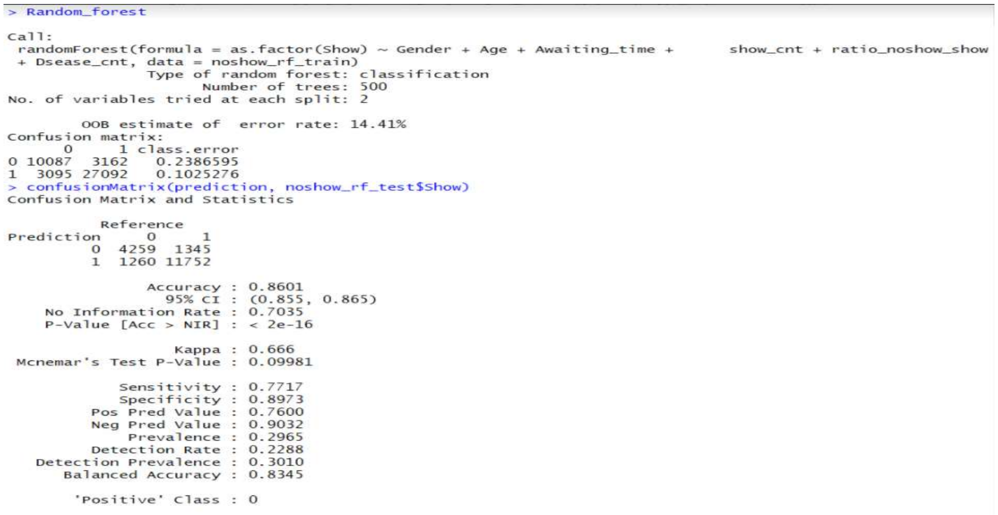
Output



Model 2 - Random Forest

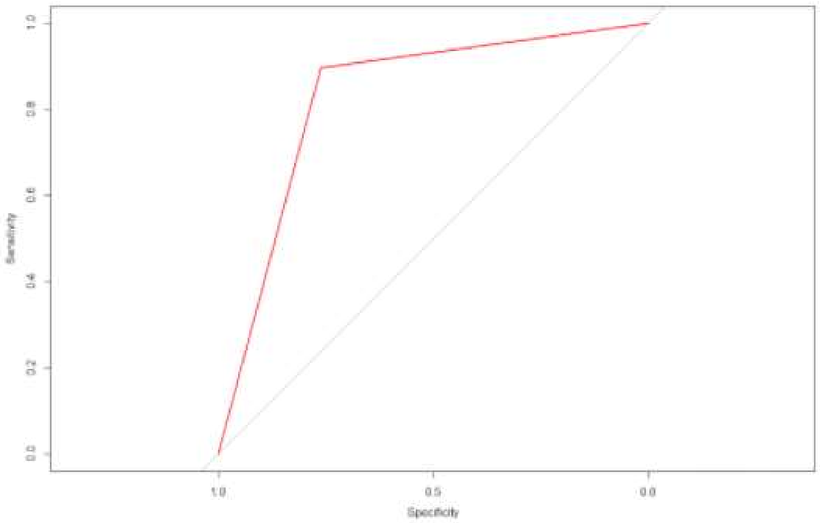
Random forests is a notion of the general technique of random decision forests that are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees habit of overfitting to their training set

Random Forest code in R

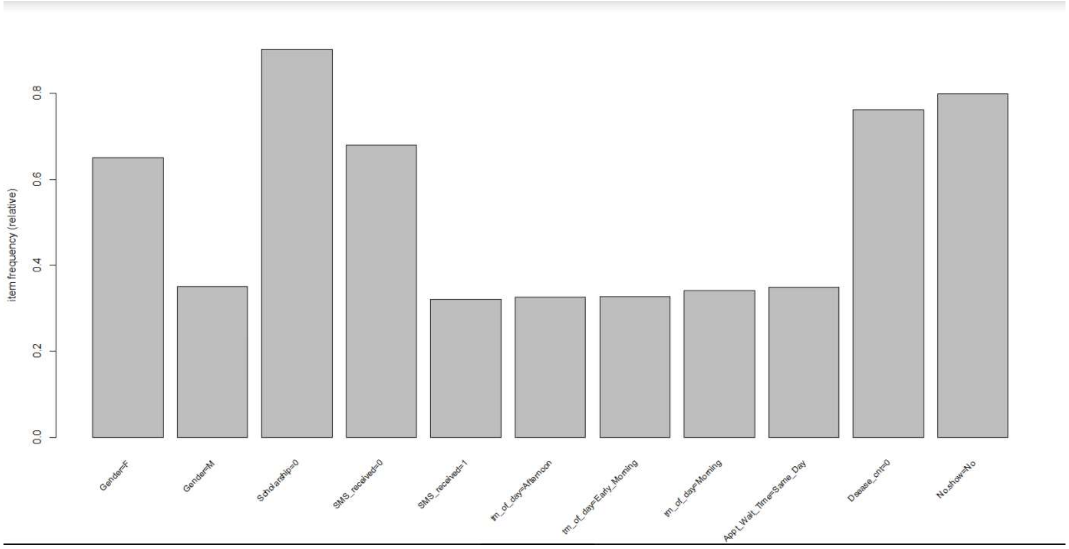


```
> Random_forest

Call:
 randomForest(formula = as.factor(Show) ~ Gender + Age + Awaiting_time +      show_cnt + ratio_noshow_show
 + Dsease_cnt, data = noshow_rf_train)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 14.41%
Confusion matrix:
      0     1 class.error
0 10087  3162   0.2386595
1  3095 27092   0.1025276
> confusionMatrix(prediction, noshow_rf_test$Show)
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0  4259  1345
         1  1260 11752

               Accuracy : 0.8601
                 95% CI : (0.855, 0.865)
    No Information Rate : 0.7035
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.666
 Mcnemar's Test P-Value : 0.09981

            Sensitivity : 0.7717
            Specificity : 0.8973
         Pos Pred Value : 0.7600
         Neg Pred Value : 0.9032
             Prevalence : 0.2965
         Detection Rate : 0.2288
   Detection Prevalence : 0.3010
      Balanced Accuracy : 0.8345

       'Positive' Class : 0
```

Output:

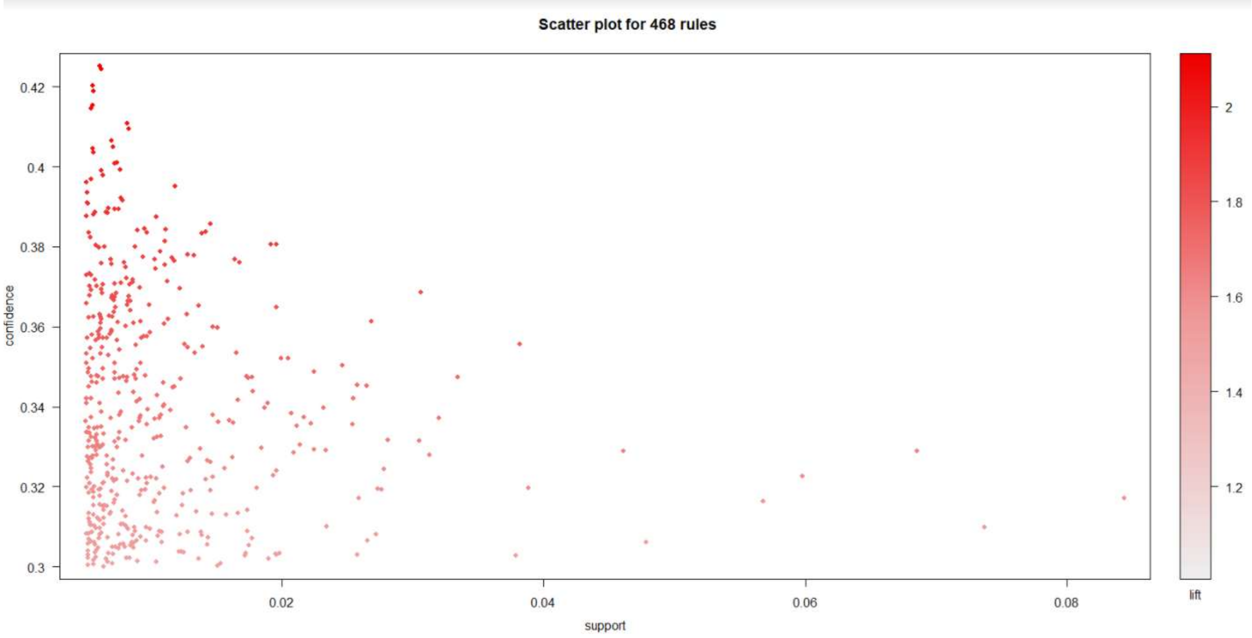


Model 3 - Association Analysis

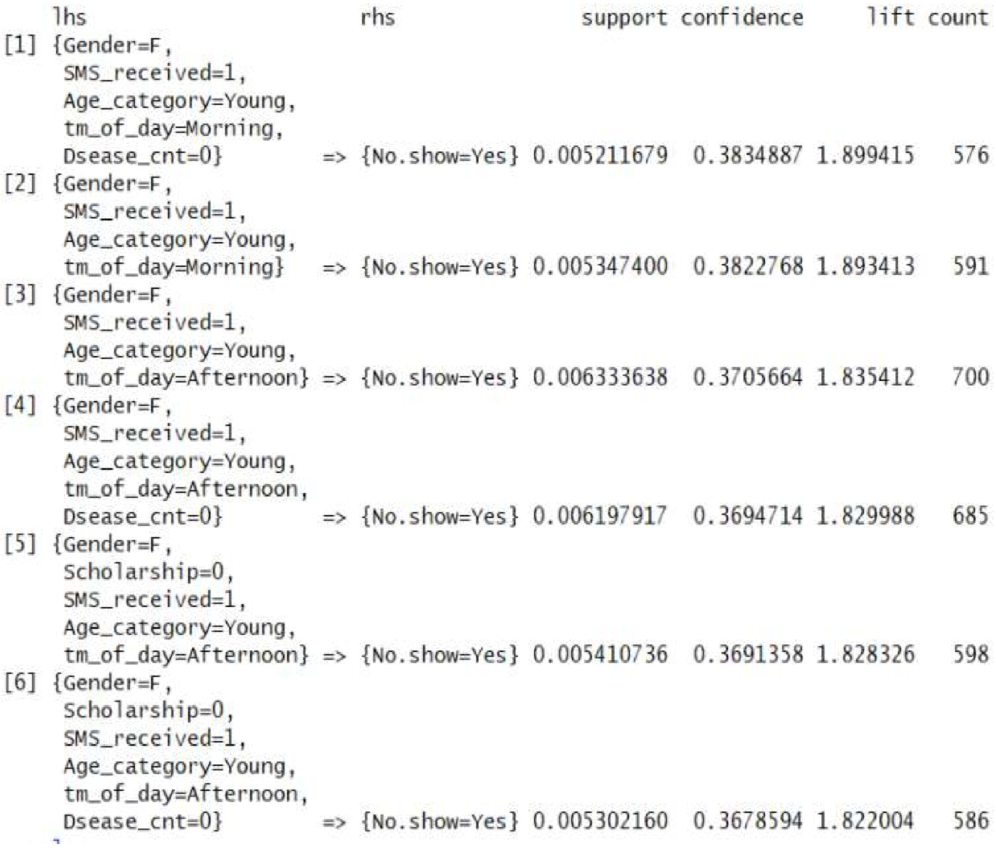


Association Analysis (Model 1)

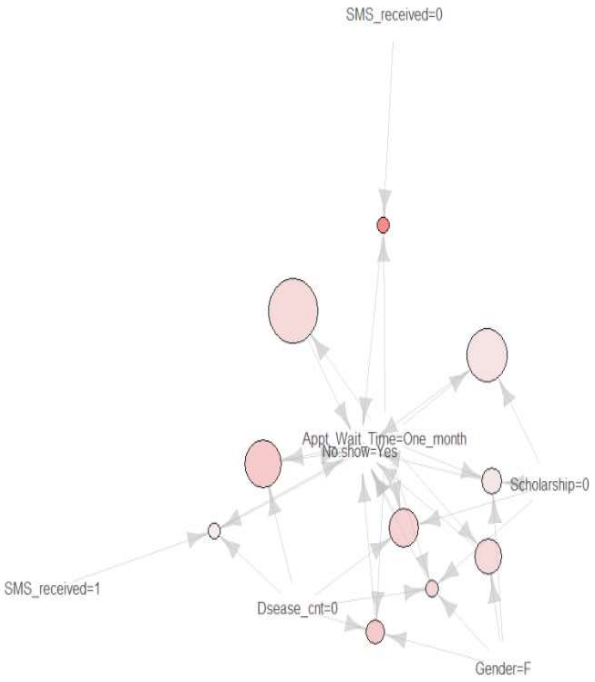
All 9 Attributes taken: Gender, Scholarship, SMS, Appt Day, Age, Time, Wait time, Disease count, No_show
Min Length: 1, Max Length: 15
Output: 468 rules



Association Analysis (Model 1 contd. Top 10 rules)

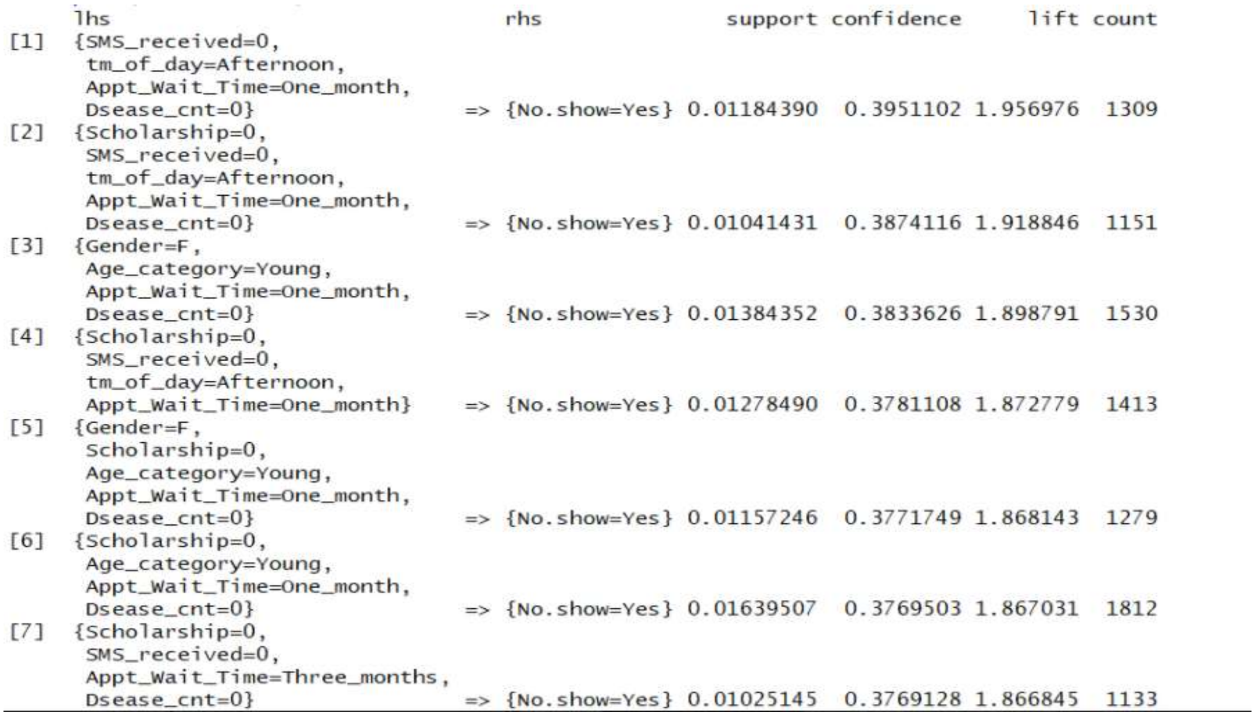


```
    lhs                          rhs               support confidence     lift count
[1] {Gender=F,
     SMS_received=1,
     Age_category=Young,
     tm_of_day=Morning,
     Dsease_cnt=0}            => {No.show=Yes} 0.005211679  0.3834887 1.899415   576
[2] {Gender=F,
     SMS_received=1,
     Age_category=Young,
     tm_of_day=Morning}       => {No.show=Yes} 0.005347400  0.3822768 1.893413   591
[3] {Gender=F,
     SMS_received=1,
     Age_category=Young,
     tm_of_day=Afternoon}     => {No.show=Yes} 0.006333638  0.3705664 1.835412   700
[4] {Gender=F,
     SMS_received=1,
     Age_category=Young,
     tm_of_day=Afternoon,
     Dsease_cnt=0}            => {No.show=Yes} 0.006197917  0.3694714 1.829988   685
[5] {Gender=F,
     Scholarship=0,
     SMS_received=1,
     Age_category=Young,
     tm_of_day=Afternoon}     => {No.show=Yes} 0.005410736  0.3691358 1.828326   598
[6] {Gender=F,
     Scholarship=0,
     SMS_received=1,
     Age_category=Young,
     tm_of_day=Afternoon,
     Dsease_cnt=0}            => {No.show=Yes} 0.005302160  0.3678594 1.822004   586
```



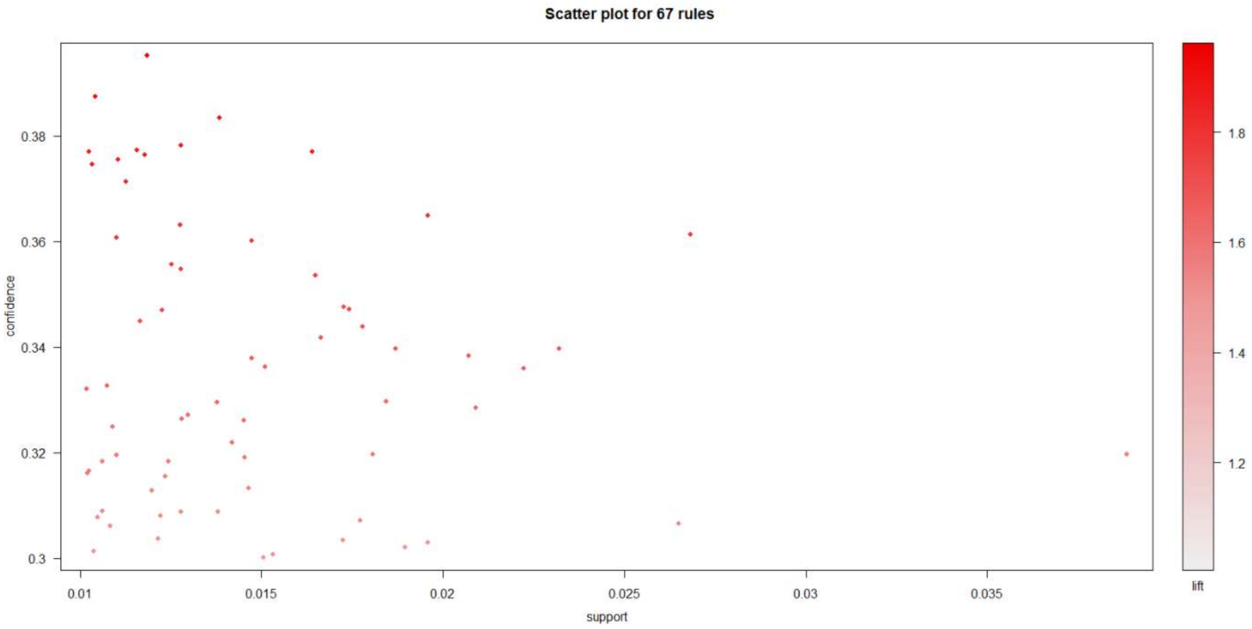
Association Model 2

All 9 Attributes taken, Min Length: 1, Max Length:15, Output: 67 rules



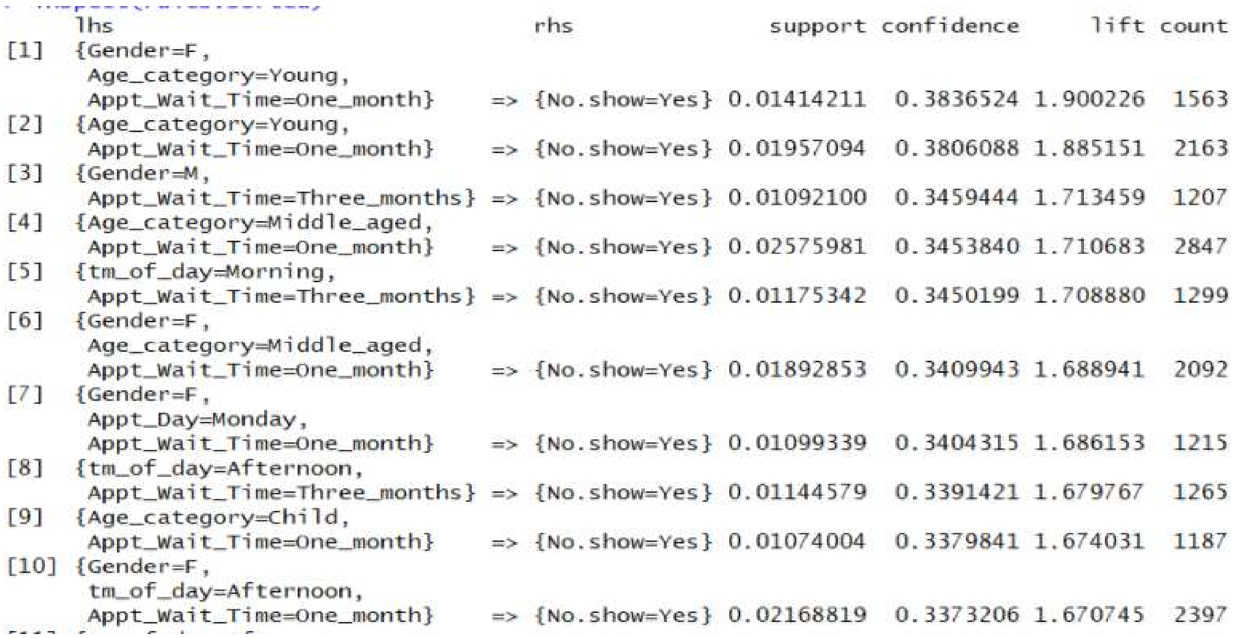
```
      lhs                              rhs             support    confidence  lift     count
[1]   {SMS_received=0,
       tm_of_day=Afternoon,
       Appt_Wait_Time=One_month,
       Dsease_cnt=0}                => {No.show=Yes} 0.01184390  0.3951102 1.956976  1309
[2]   {Scholarship=0,
       SMS_received=0,
       tm_of_day=Afternoon,
       Appt_Wait_Time=One_month,
       Dsease_cnt=0}                => {No.show=Yes} 0.01041431  0.3874116 1.918846  1151
[3]   {Gender=F,
       Age_category=Young,
       Appt_Wait_Time=One_month,
       Dsease_cnt=0}                => {No.show=Yes} 0.01384352  0.3833626 1.898791  1530
[4]   {Scholarship=0,
       SMS_received=0,
       tm_of_day=Afternoon,
       Appt_Wait_Time=One_month}    => {No.show=Yes} 0.01278490  0.3781108 1.872779  1413
[5]   {Gender=F,
       Scholarship=0,
       Age_category=Young,
       Appt_Wait_Time=One_month,
       Dsease_cnt=0}                => {No.show=Yes} 0.01157246  0.3771749 1.868143  1279
[6]   {Scholarship=0,
       Age_category=Young,
       Appt_Wait_Time=One_month,
       Dsease_cnt=0}                => {No.show=Yes} 0.01639507  0.3769503 1.867031  1812
[7]   {Scholarship=0,
       SMS_received=0,
       Appt_Wait_Time=Three_months,
       Dsease_cnt=0}                => {No.show=Yes} 0.01025145  0.3769128 1.866845  1133
```

Lift : 1.96, Support : 0.012, Confidence : 0.39



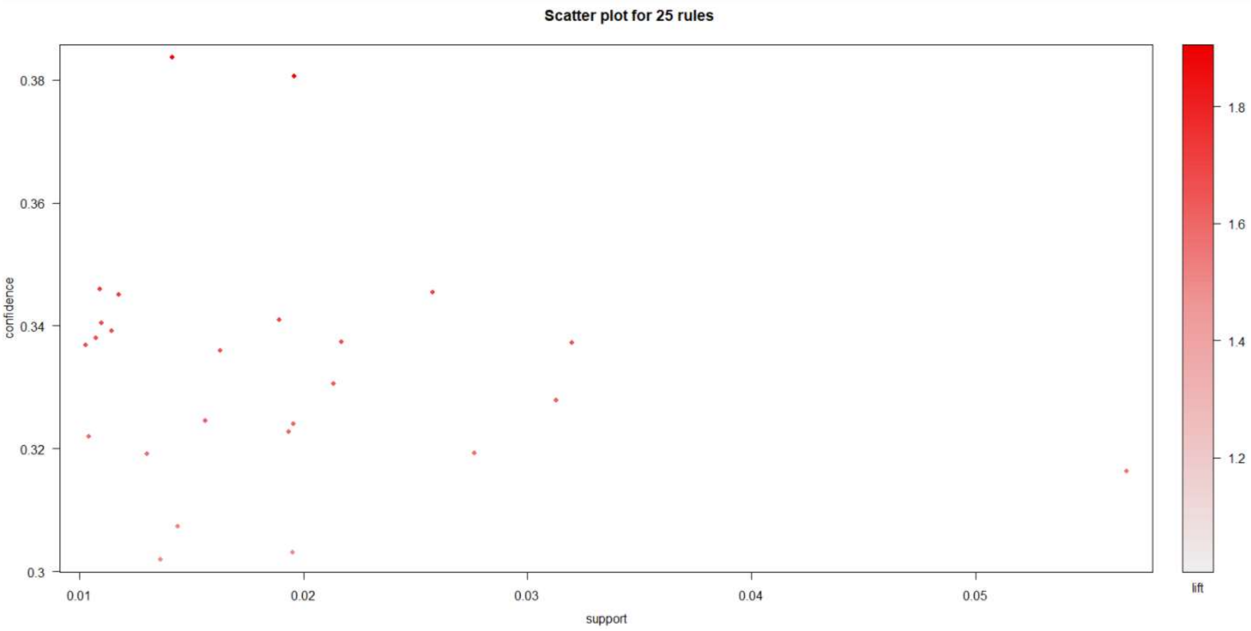
Association Model 3

Attributes taken: Gender, appt_Day, Age_category, tm_of_day, Appt_Wait_Time, No.show
Min Length: 3, Max Length:12, Output: 25 rules



```
      lhs                                rhs             support confidence     lift count
[1]   {Gender=F,
       Age_category=Young,
       Appt_Wait_Time=One_month}     => {No.show=Yes} 0.01414211  0.3836524 1.900226  1563
[2]   {Age_category=Young,
       Appt_Wait_Time=One_month}     => {No.show=Yes} 0.01957094  0.3806088 1.885151  2163
[3]   {Gender=M,
       Appt_Wait_Time=Three_months} => {No.show=Yes} 0.01092100  0.3459444 1.713459  1207
[4]   {Age_category=Middle_aged,
       Appt_Wait_Time=One_month}     => {No.show=Yes} 0.02575981  0.3453840 1.710683  2847
[5]   {tm_of_day=Morning,
       Appt_Wait_Time=Three_months} => {No.show=Yes} 0.01175342  0.3450199 1.708880  1299
[6]   {Gender=F,
       Age_category=Middle_aged,
       Appt_Wait_Time=One_month}     => {No.show=Yes} 0.01892853  0.3409943 1.688941  2092
[7]   {Gender=F,
       Appt_Day=Monday,
       Appt_Wait_Time=One_month}     => {No.show=Yes} 0.01099339  0.3404315 1.686153  1215
[8]   {tm_of_day=Afternoon,
       Appt_Wait_Time=Three_months} => {No.show=Yes} 0.01144579  0.3391421 1.679767  1265
[9]   {Age_category=Child,
       Appt_Wait_Time=One_month}     => {No.show=Yes} 0.01074004  0.3379841 1.674031  1187
[10]  {Gender=F,
       tm_of_day=Afternoon,
       Appt_Wait_Time=One_month}     => {No.show=Yes} 0.02168819  0.3373206 1.670745  2397
```

Lift : 1.9, Support : 0.014, Confidence : 0.384



Conclusions/Recommendation

After analyzing all the models above, we came to below conclusions/recommendations.

Conclusions

- After Association Analysis, we found that Disease_count, Young age category & Awaiting time are strong appointment status determinant.
- Random Forest model performs best. Association analysis complements our results of random forest.

Recommendation for Stakeholders

- Doctors : Based our model, doctors can do efficient scheduling.
- Health Centers : No Investment in SMS is needed.

References