**MINI-PROJECT-REPORT**

**FERTILITY\_DIAGNOSIS**

**MINI-PROJECT-REPORT**

**COURSE CODE/TITLE:17fe32/17le32/Analytics with R Programming Alongside with Visualization**

Year:III sem:V

17F231-Preethi M-Information Technology

17L307-Bhargavi A-Electronics and Communication Engineering

**DESCRIPTION ABOUT PROJECT:**

This dataset contains the data given by 100 volunteers that provides a semen sample analyzed according to the WHO 2010 criteria. Sperm concentration are related to socio-demographic data, environmental factors, health status, and life habits .This dataset gives the fertility analysis based on some attributes ,by visualizing those data we can predict the outcome of a testing based on fertility.

DESCRIPTION OF ATTRIBUTES:

This consists of **10** attributes:

**1st attribute** :Season in which the analysis was performed. 1) winter, 2) spring, 3) Summer, 4) fall. (-1, -0.33, 0.33, 1)

**2nd attribute:** Age at the time of analysis. 18-36 (0, 1)

**3rd attribute:** Childish diseases (ie , chicken pox, measles, mumps, polio) 1) yes, 2) no. (0, 1)

**4th attribute:** Accident or serious trauma 1) yes, 2) no. (0, 1)

**5th attribute:** Surgical intervention 1) yes, 2) no. (0, 1)

**6th attribute:** High fevers in the last year 1) less than three months ago, 2) more than three months ago, 3) no. (-1, 0, 1)

**7th attribute:** Frequency of alcohol consumption 1) several times a day, 2) every day, 3) several times a week, 4) once a week, 5) hardly ever or never (0, 1)

**8th attribute:** Smoking habit 1) never, 2) occasional 3) daily. (-1, 0, 1)

**9th attribute**: Number of hours spent sitting per day ene-16 (0, 1)

**10th  attribute:** Output: Diagnosis normal (N), altered (O)

**Machine Learing Steps:**

* Loading a file-using delim
* Viewing a file-in the form a table
* Naming the columns-such that each data can be accessed easliy
* Preparing the data for analyzing-mutate is used to modify the given data for visualizing
* Shuffling the data-this is used to get the accurate results
* Cleaning the data-such that no null values can enter into the visualizing
* Separating the data into test set and training data sets
* Visualizing the data in the form of decision tree
* Testing the accuracy

**Diagram:**

Loading a file

Viewing a file

Preparing the data for analyzing

Cleaning the data

Separating the data into test set and training data sets

Visualizing the data in form of decision tree

Testing the accuracy

# Exploratory Data Analysis:

# Library files that needed to be import:

# library(tidyr)

# library(dplyr)

# library(ggplot2)

library(plotly)

# Script And Ouput:

# dataset=read.delim("fertility\_Diagnosis.txt",sep=",",header=F)

# View(dataset)

# 

# colnames(dataset)=c("seasons","age","childish\_diseases","accidents\_trauma", "surgical\_intervention","high\_fevers","alcohol","smoking","no\_of\_hours","output")

# View(dataset)

# 

# head(dataset)

# 

# tail(dataset)

# 

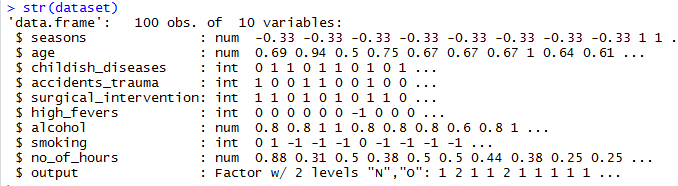
# levels(dataset$output)

# 

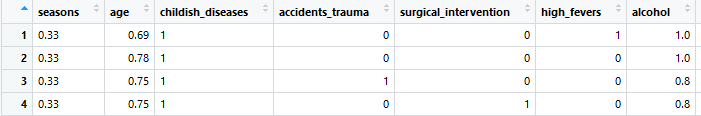
# glimpse(dataset)

# 

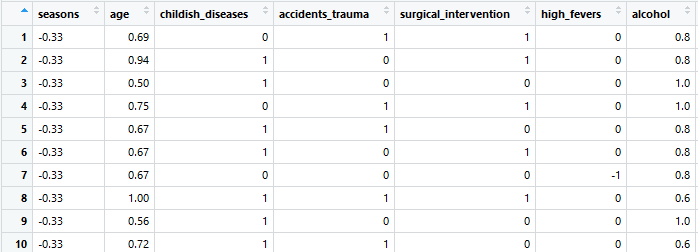
# str(dataset)

****

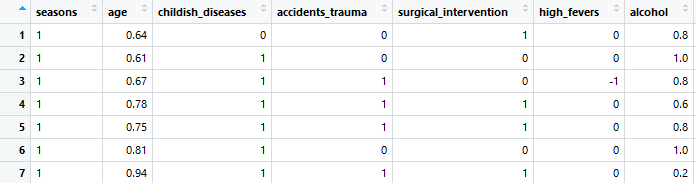
View(dataset %>% filter(dataset$seasons==0.33))



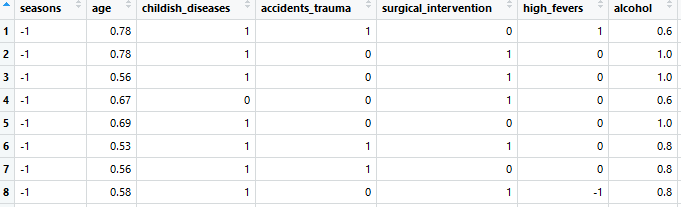
View(dataset %>% filter(dataset$seasons==-0.33))



View(dataset %>% filter(dataset$seasons==1))

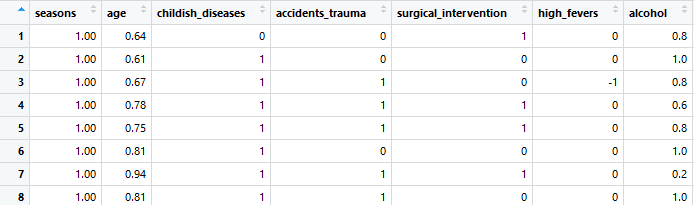


View(dataset %>% filter(dataset$seasons==-1))



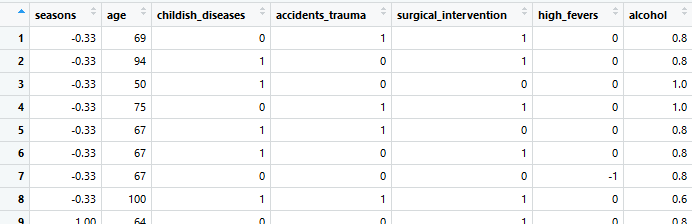
#arrange

View(dataset %>% arrange(desc(dataset$seasons)))



#mutate

View(dataset %>% mutate(age=age\*100))

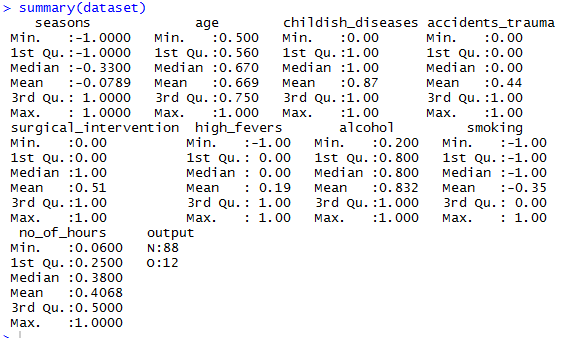


#group\_by

#View(group\_by(dataset,seasons))

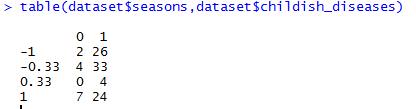
#summarize

summary(dataset)



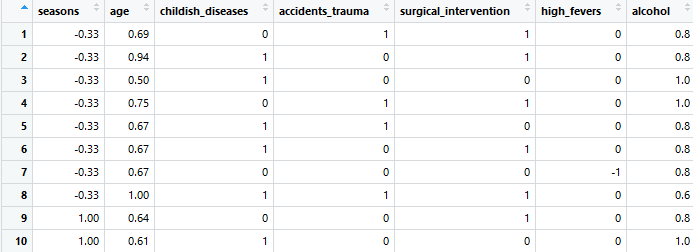
#contingency table

table(dataset$seasons,dataset$childish\_diseases)



#drop levels:

View(dataset %>% filter(seasons!=-1) %>% droplevels())

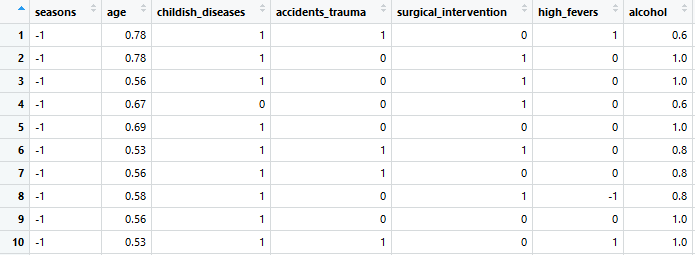


#visualization

library(ggplot2)

data\_winter=dataset %>% filter(dataset$seasons==-1)

View(data\_winter)

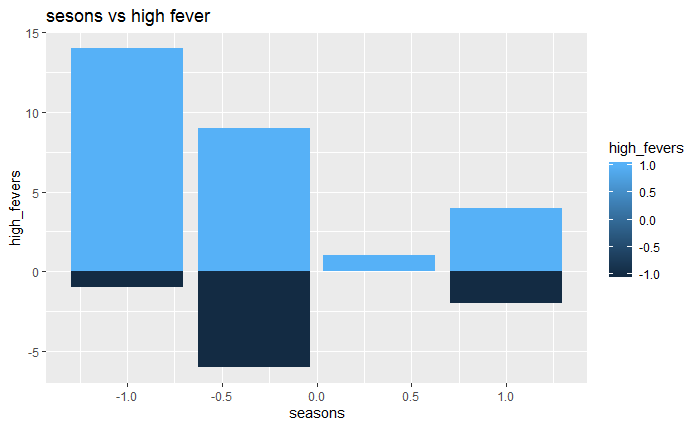


install.packages("plotly")

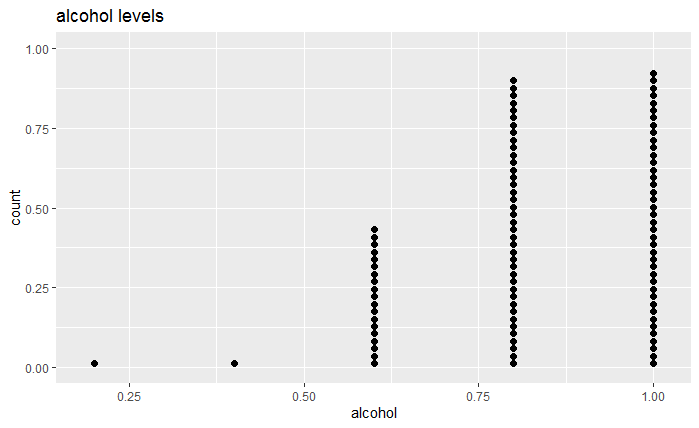
library(plotly)

#visualizing the fever rate based on seasons

ggplot(data=dataset,aes(x=seasons,y=high\_fevers,fill=high\_fevers))+geom\_col()+ggtitle("sesons vs high fever")



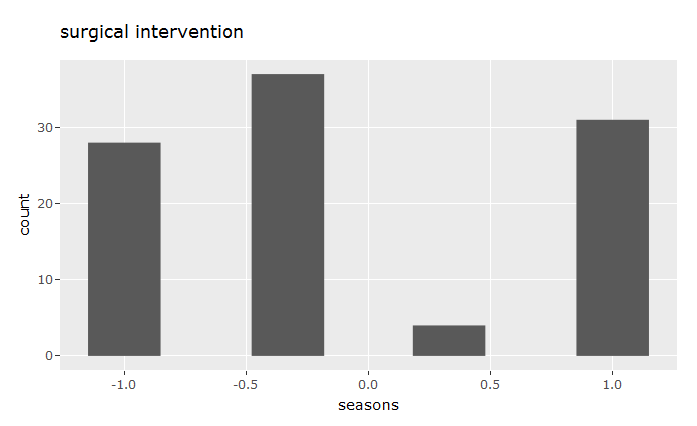
ggplot(data=dataset,aes(x=alcohol))+geom\_dotplot(dotsize=0.4)+ggtitle("alcohol levels")



#ggplot using plotly

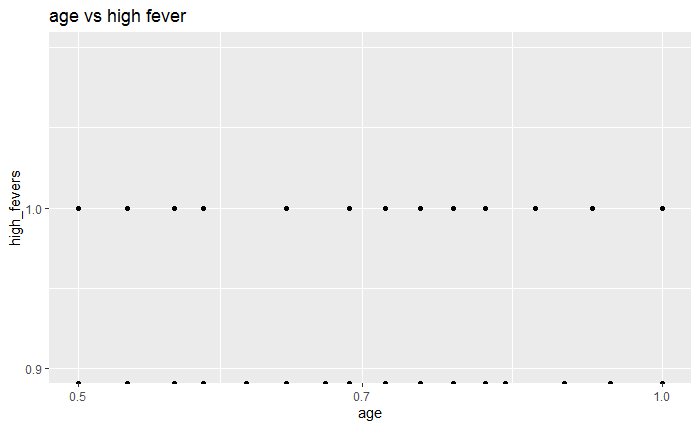
a=ggplot(data=dataset,aes(x=seasons,fill=surgical\_intervention,color=surgical\_intervention))+geom\_bar()+ggtitle("surgical intervention")

ggplotly(a)

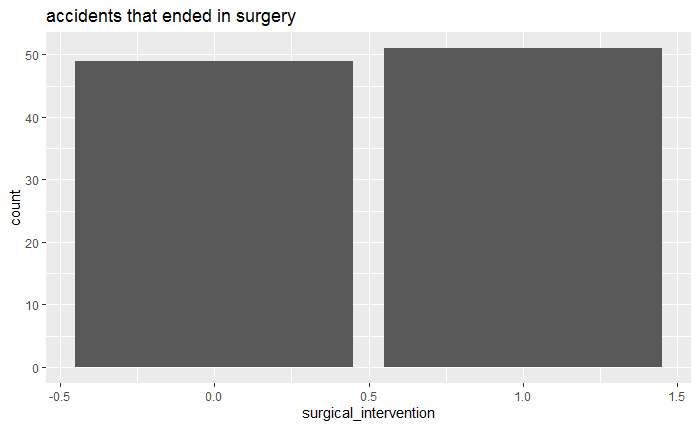


#dot plot

ggplot(data=dataset,aes(x=age,y=high\_fevers))+geom\_point()+scale\_x\_log10()+scale\_y\_log10()+ggtitle("age vs high fever")

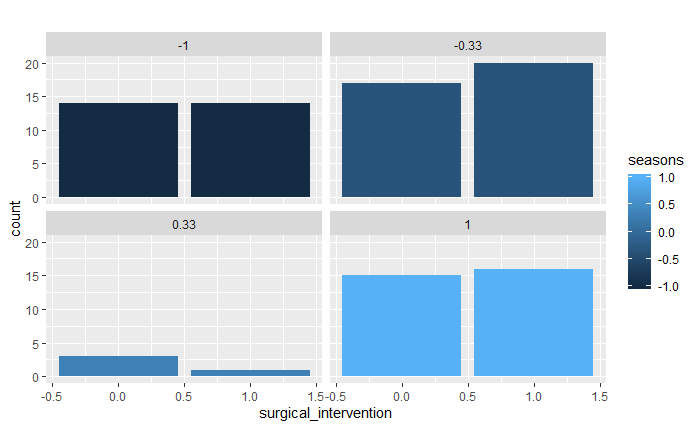
#Bar plot

ggplot(data=dataset,aes(x=surgical\_intervention,fill=accidents\_trauma))+ geom\_bar(position = "dodge")+ggtitle("accidents that ended in surgery")



ggplot(data=dataset,aes(x=surgical\_intervention,fill=seasons,color=surgical\_intervention))+

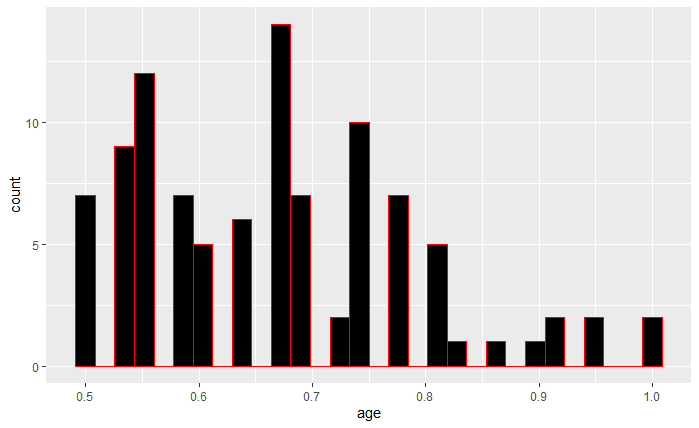
geom\_bar()+facet\_wrap(~seasons)+ggtitle("")



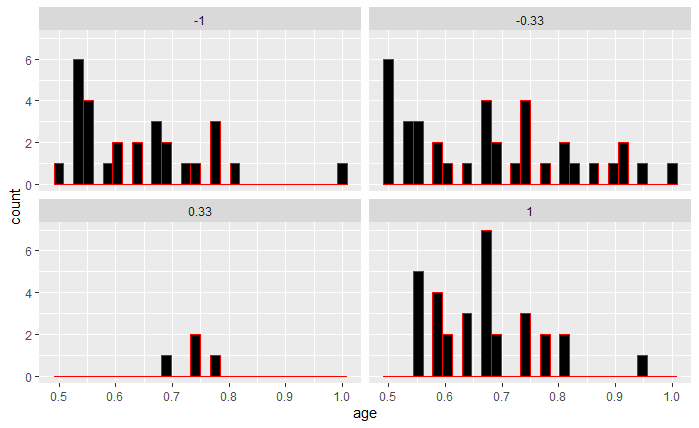
#Histogram

#ggplot(data=dataset,aes(x=age))+geom\_histogram()

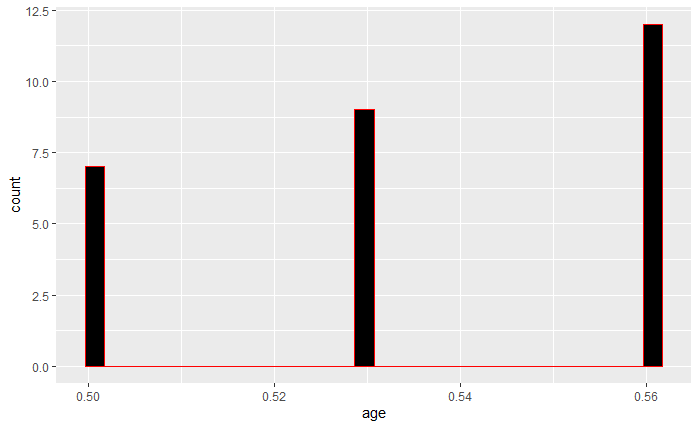
ggplot(data=dataset,aes(x=age))+geom\_histogram(col="red",fill="black")



ggplot(data=dataset,aes(x=age))+geom\_histogram(col="red",fill="black")+facet\_wrap(~seasons)

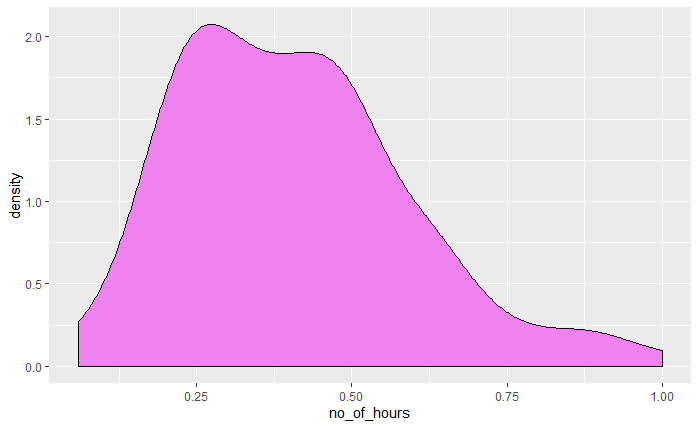


dataset %>% filter(age<0.58) %>% ggplot(aes(x=age))+geom\_histogram(col="red",fill="black")



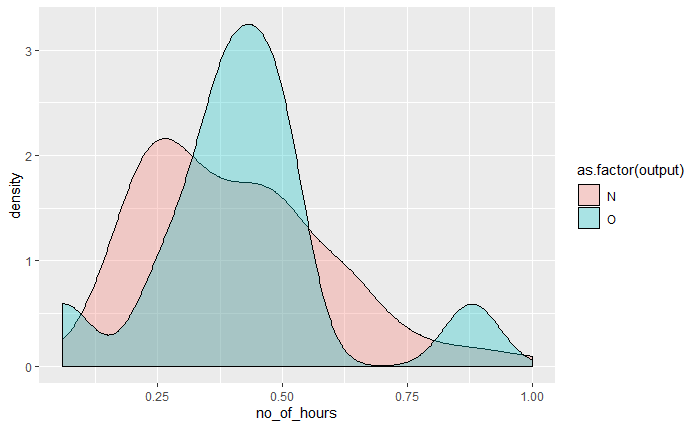
#Density plot

ggplot(data=dataset,aes(x=no\_of\_hours,fill=childish\_diseases))+geom\_density(col="black",fill="violet")



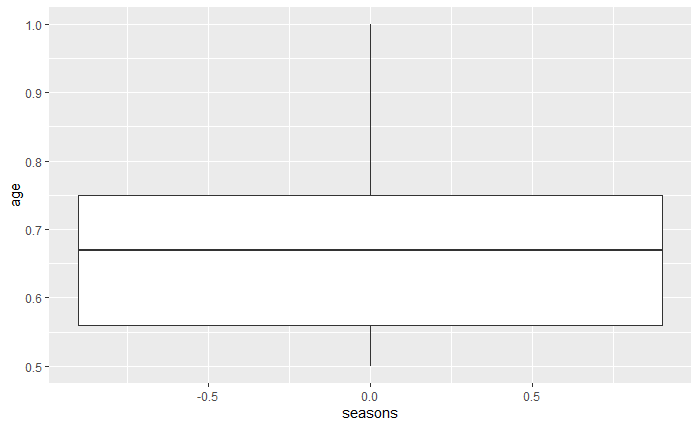
ggplot(data=dataset,aes(x=no\_of\_hours,fill=as.factor(output)))+

geom\_density(alpha=0.3)

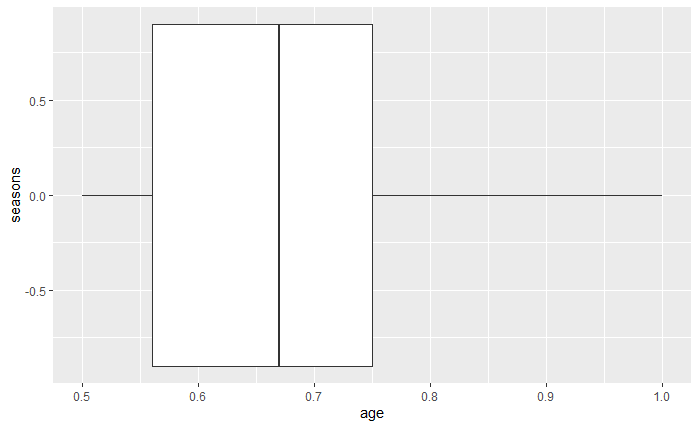


#Box plot

ggplot(data=dataset,aes(x=seasons,y=age,group=1))+geom\_boxplot()

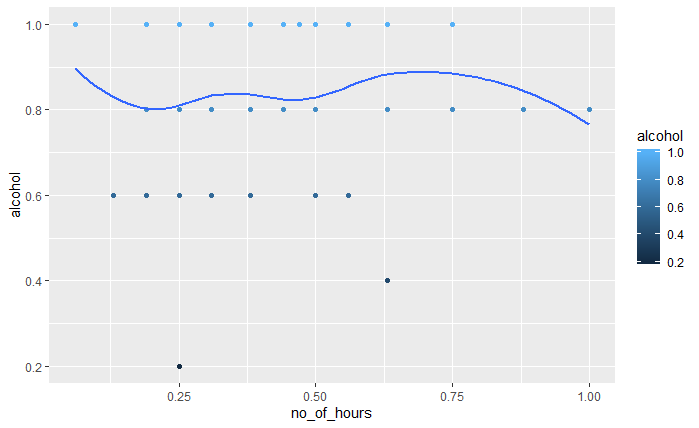


ggplot(data=dataset,aes(x=seasons,y=age,group=1))+geom\_boxplot()+coord\_flip()



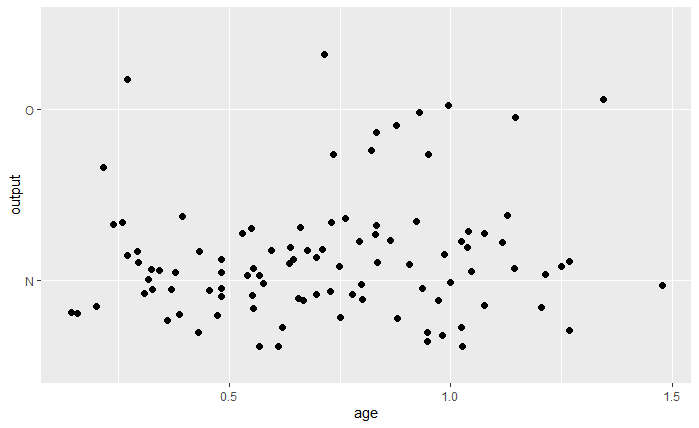
#point

ggplot(dataset,aes(x=no\_of\_hours,y=alcohol))+geom\_point(aes(col=alcohol))+geom\_smooth(method = "loess",se=F)



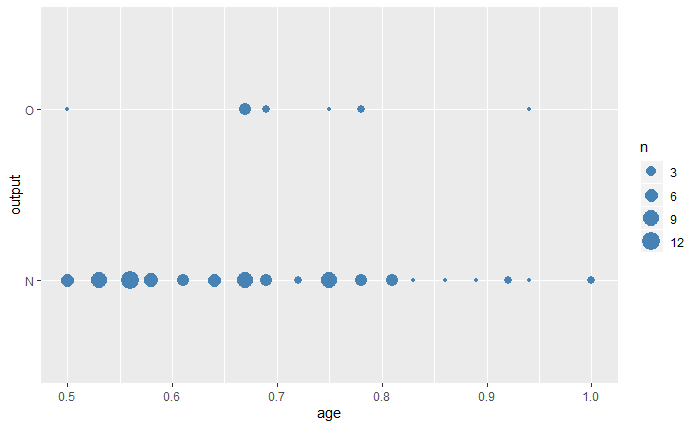
#jitter

ggplot(dataset,aes(x=age,y=output))+geom\_jitter(width=0.5,size=2)



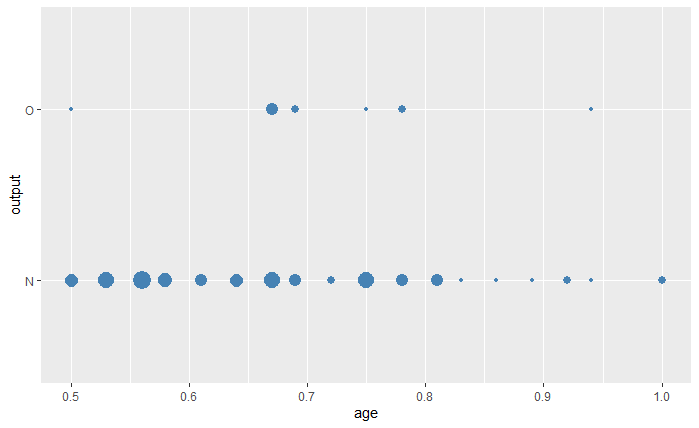
#ggcount(with legend)

ggplot(dataset,aes(x=age,y=output))+geom\_count(col="steelblue")



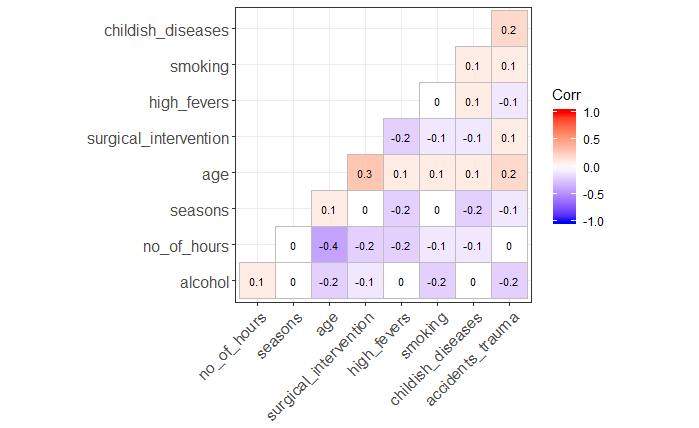
#without legend

ggplot(dataset,aes(x=age,y=output))+geom\_count(col="steelblue",show.legend=F)



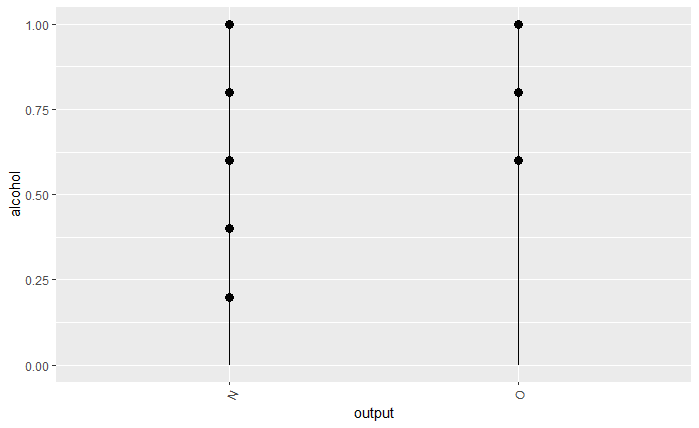
corr <- round(cor(dataset[,-10]),1)library(ggcorrplot)

ggcorrplot(corr, hc.order = TRUE, type = "lower", lab = TRUE,lab\_size = 3ggtheme=theme\_bw)



ggplot(dataset, aes(x=output, y=alcohol)) +geom\_point(size=3) + geom\_segment(aes(x=output,

xend=output, y=0, yend=alcohol))+theme(axis.text.x = element\_text(angle=65, vjust=0.6))



**Machine learning:**

#machine learning concepts

#logistic regression and decesion tree

dataset=read.delim("fertility\_Diagnosis.txt",sep=",",header=F)

View(dataset)

colnames(dataset)=c("seasons","age","childish\_diseases","accidents\_trauma",

"surgical\_intervention","high\_fevers","alcohol","smoking",

"no\_of\_hours","output")

View(dataset)

library(dplyr)

dataset=dataset %>% mutate(age=age\*100)

dataset=dataset %>% mutate(no\_of\_hours=((no\_of\_hours\*100)\*24)/120)

dataset=dataset %>% mutate(alcohol=alcohol\*100)

View(dataset)

#age,trauma,surgery,sitting

#since there is no linearity between the data we go for decision tree

set.seed(678)

shuffle\_index=sample(1:nrow(dataset))

dataset=dataset[shuffle\_index,]

dim(dataset)

clean\_data=dataset %>%select(c("seasons","childish\_diseases","high\_fevers","alcohol","smoking"))

str(clean\_data)

library(caTools)

split=sample.split(clean\_data$output,SplitRatio=0.8)

training\_set=subset(clean\_data,split==TRUE)

test\_set=subset(clean\_data,split==FALSE)

View(training\_set)

View(test\_set)

#build a model

library(rpart)

str(clean\_data)

fit=rpart(formula=output~.,data=training\_set,method="class")

library(rpart.plot)

rpart.plot(fit)

str(clean\_data)

library(plotrix)

pred\_unseen=predict(object=fit,newdata=test\_set,type="class")

pred\_unseen

table\_mat=table(test\_set$output,pred\_unseen)

table\_mat[1,1]

table\_mat

accuracy=sum(diag(table\_mat))/sum(table\_mat)

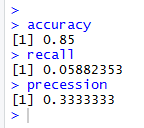
accuracy

recall=table\_mat[2,2]/(sum(diag(table\_mat)))

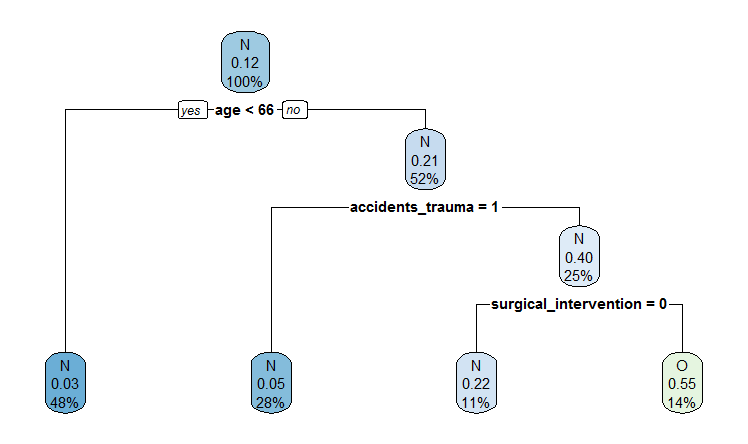
recall

precession=table\_mat[2,2]/(table\_mat[2,2]+table\_mat[1,2])

precession



**Decision tree output:**



**Conclusion:**

Thus we have analysed the given data set successfully.

**Reference:**

Rpubs:<http://rpubs.com/Preethi_17f231/fertility_analysis>