TAREA M3

TATAN RUFINO

10 de mayo de 2019

1. INTRODUCCIÓN

Con este proyecto trataré de hacer un modesto análisis de datos sobre la base de datos student proporcionada por la comunidad internacional y sobre la que se centra la tarea en la que nos encontramos. Me he ayudado de la comunidad internacional, que ofrece todo tipo de ayuda en análisis para bases de datos conocidas, razón por la cual usaré funciones y metologías no mostradas hasta el momento

Este proyecto tiene como objetivo descubrir la influencia de diversas variables y parámetros en las notas de un estudiante de secundaria. En este caso nos centraremos en el estudiante de matemáticas, las observaciones fueron recabadas haciendo la encuesta de un total de 395 estudiantes, que viene dado por el dataset studentMat.csv.

El dataset cuenta con 33 variables diferentes, el cual será sometido a la siguiente metodología de acciones: - INICIO: Declaración de librerías, importación de tablas, primeras visualizaciones - Filtrado de dataset y definición de variables clave - Análisis exploratorio y visualizaciones - Comentarios - Aplicación de algoritmos de Machine learning - Algoritmo de Arbol de decisiones - Algoritmo de Regresión simple - Evaluación

2. INICIO

```
temp <- tempfile()
download.file("http://archive.ics.uci.edu/ml/machine-learning-databases/00356/student.zip",temp, mode="wb")
unzip(temp, "student-mat.csv")
ST <- read.table("student-mat.csv",sep= ";", header= T)
unlink(temp)

ST[,1:5] %>% head() %>% kable()

summary(head(ST[,1:5])
```

3. FILTRADO Y DEFINICION DE VARIABLES CLAVE

```
temp <- tempfile()
download.file("http://archive.ics.uci.edu/ml/machine-learning-databases/00356/student.zip",temp, mode="wb")
unzip(temp, "student-mat.csv")
ST <- read.table("student-mat.csv",sep= ";", header= T)
unlink(temp)
```

Primeras visualizaciones

```
data[,1:5] %>% head() %>% kable()
summary(ST)
names(ST)
```

Compruebo si hay NA

```
ST %>% is.na() %>% all()
```

Consulto la codificacion de la DT de la pagina de google

```
COL_LIST <- gsheet2tbl('https://docs.google.com/spreadsheets/d/1mDsF0aMNgODx706312mxV1_zP32fAe_P55SBmtG72G8'
)
kable(COL_LIST)
kable(COL_LIST[,1:3])
```

Escogemos 18 variables como predictores - 7 variables categóricas - 11 variables numericas - Y finalmente 3 variables dependientes y susceptibles de predicción, correspondientes a las notas

En base a lo que observo en nuestro ST, y dadas las características de las variables actuales, eliminaré diversas variables entre las que se encuentran “famsize” “reason” o “guardian” o “school”, ya que serán irrelevantes en nuestro análisis con select()

```
ST <- ST %>%as_tibble()%>%select(sex, age, address,Pstatus, Medu, Fedu, Mjob, Fjob,studytime,traveltime,fail
ures,higher,internet,
         goout, Dalc,Walc,health, absences,G1,G2,G3)
dim(ST)

ST %>% glimpse()
ST %>% summary()
```

4. ANÁLISIS EXPLORATORIO DE DATOS DIVERSO

Analizaremos los siguientes indicadores

- 1. Según género
- 2. Consumo de alcohol (W/D)
- 3. Costumbres, objetivos y origen del alumno
- 4. Salud y asistencia a clase
- 5. Acceso a internet
- 6. Relación asistencia a clase con las notas

Según Genero

```
(ANA_1<-ST%>%
    mutate(pass=ifelse(G3>=10,1,0), fail= ifelse(G3<10,1,0))%>%
    filter(sex=="F"|sex=="M")%>%
    group_by(sex)%>%
    summarise(Pass=sum(pass),
              Fail=sum(fail)))

ANA_1%>%
  ggplot(aes(x=sex,y=Fail))+
  geom_bar(stat="identity")
ANA_1
```

Observamos que hay diferencias notables entre mujeres y hombres

4.2. Consumo de alcohol

```
ANA_2a <- ST%>%
  group_by(Walc)%>%
  aggregate(G3~Walc, data=., mean)%>%
  arrange(desc(G3))
ANA_2a

ANA_2b <- ST%>%
  group_by(Dalc)%>%
  aggregate(G3~Dalc, data=., mean)%>%
  arrange(desc(G3))
ANA_2b
```

Vemos que disminuye el promedio a medida que aumenta el consumo de alcohol entre semana, pero no hay un patron para fin de semana

Cruce de datos para visualizacion de aspectos que relacionen datos relacionados con el género y el consumo de alcohol

```
DUM <- dummyVars("~.", data=ST)
GEN_ALC <- data.frame(predict(DUM, newdata=ST))
correl1 <-cor(GEN_ALC[,c("G3","sex.F","sex.M","Walc","Dalc")])
source("https://raw.githubusercontent.com/briatte/ggcorr/master/ggcorr.R")
correl1 %>%
  ggcorr(label = TRUE)+ ggtitle("Correlaciones entre genero, consumo de alcohol y las notas")

ST$Dalc <- as.factor(ST$Dalc)
ST$Walc <- as.factor(ST$Walc)
P1a<-ST %>%
  ggplot(aes(x=Dalc, y=G3, fill= Dalc))+
  geom_boxplot()+
  coord_flip()+
  xlab("Consumo de alcohol entre semana")+
  ylab("Notas")+
  facet_grid(~sex)
P1b<-ST %>%
  ggplot(aes(x=Walc, y=G3, fill= Walc))+
  geom_boxplot()+
  coord_flip()+
  xlab("Consumo de alcohol entresemana")+
  ylab("Notas")+
  facet_grid(~sex)
grid.arrange(P1a,P1b,ncol=2)
```

Básicamente podemos predecir que el consumo de alcohol tiene un impacto mucho mayor que el género en las notas

4.3. Costumbres, objetivos y origen del alumno

```
class(ST$goout)
ST$goout <- as.factor(ST$goout)

>Convertimos a factor para que me acepte el algoritmo

ANA_3 <- ST%>%
  group_by(goout)%>%
  summarise(Avr= mean(G3,na.rm=TRUE))%>%
  arrange(desc(Avr))
ANA_3
```

Vemos que hay un decrecimiento en las notas en cuanto que el alumno supera el factor de salida de 4

```
P2a<-ST %>%
  group_by(address)%>%
  ggplot(aes(x=factor(Dalc), y= G3))+
  geom_jitter(alpha=0.6)+
  scale_x_discrete("Alcohol entresemana")+
  scale_y_continuous("Notas")+
  facet_grid(~address)
P2b<-ST %>%
  group_by(address)%>%
  ggplot(aes(x=factor(Walc), y= G3))+
  geom_jitter(alpha=0.6)+
  scale_x_discrete("Alcohol en fin de semana")+
  scale_y_continuous("Notas")+
  facet_grid(~address)
grid.arrange(P2a,P2b,ncol=2)
```

Otro ejemplo paradigmático de la relacion del decrecimiento de notas respecto al consumo de alcohol. Se observa que se bebe menos en las zonas rurales

```
ST%>%
  ggplot(aes(x=higher, y=G3))+
  geom_boxplot()+
  facet_grid(~sex)
```

Los alumnos que aspiran a mejor educacion, tienen mejores notas. Los hombres son mejores que las mujeres también

4.4. Salud y asistencia a clase

```
ST%>%
  group_by(sex)%>%
  ggplot(aes(x=factor(health), y=absences, color=sex))+
  geom_smooth(aes(group=sex), method="lm", se=FALSE)
```

Este interesante gráfico nos muestra que existe una relación lineal decreciente entre la salud del alumno y las ausencias en clase Ademas de ser, de nuevo, las mujeres quienes faltan más clase, siendo sus ausencias menos dependientes con la salud

4.5. Acceso a internet

```
ST%>%
  group_by(internet)%>%
  ggplot(aes(x=G3, fill=internet))+
  geom_density( alpha=0.8)
```

el uso de internet afecta a las notas, aunque no en exceso

4.6. Relación asistencia a clase con las notas

```
P3 <- ggplot(ST, aes(absences, G3))
P3 + geom_point() +
  geom_smooth(method="lm", se=F) +
  labs(y="G3",
       x="abcenses",
       title="Comparativa ausencias con notas")

ggplot(ST, aes(x=absences, y=G3)) +
  geom_bar(stat="identity", width=.5, fill="tomato3") +
  labs(y="G3",
       x="absences",
    title="Ausencias vs Notas")
```

5. ANALISIS NO SUPERVISADO SEGÚN CLUSTERING CON ALGORITMO KMEANS

CLUSTERING DE DATOS >CREAMOS ST_MOD COMO REFERENCIA CON EL G3 REESCALADO DE 1:10

```
ST_MOD=ST
ST_MOD$G3=as.integer(ST_MOD$G3)

ST_MOD$G3=rescale(ST_MOD$G3,to=c(1,10))

>convertimos a integer

ST_KM= ST[c("age","Medu","Fedu","studytime","traveltime","failures","goout","Dalc","Walc","health","absences
","G1","G2")]
  ST_factors = ST_KM %>% select_if(is.factor) %>% colnames()
  ST_KM[,ST_factors] = lapply(ST_KM[,ST_factors], as.integer)

  >Limpiamos variables de valores dispersos

      plot(ST_KM$absences)
      ST_KM$absences<- floor(rescale(ifelse(ST_KM$absences>3*mean(ST_KM$absences),
                                            3*mean(ST_KM$absences),ST_KM$absences),
                                     ,to=c(1,5)))

      plot(ST_KM$age)
      ST_KM$age<- floor(rescale(ifelse(ST_KM$age>19,
                                            19,ST_KM$age),
                                     ,to=c(1,5)))
      plot(ST_KM$age)

 >Se reajusta todo el sistema a escala 1:5
```

```
ST_KM[, c(1:13)] <- lapply(ST_KM[, c(1:13)], function(x) rescale(x,to=c(1,5)))
ST_KM=floor(ST_KM)
  ST_factors = ST_KM %>% select_if(is.factor) %>% colnames()
  ST_KM[,ST_factors] = lapply(ST_KM[,ST_factors], as.integer)

glimpse(ST_KM)
summary(ST_KM)

mydata <- ST_KM
wss <- (nrow(mydata)-1)*sum(apply(mydata,2,var))
for (i in 2:15) wss[i] <- sum(kmeans(mydata,centers=i)$withinss)

plot(1:15, wss, type="b", xlab="Numero de Clusters",
     ylab="Sumas de cuadrados dentro de los grupos",
     main="Num de clusters óptimo según Elbow",
     pch=20, cex=2)

>se deciden 5 centros

KM=kmeans(ST_KM,5)
KM

table(ST_MOD$G3, KM$cluster)

plot(ST_KM$age, col=KM$cluster)

plot(KM$centers)
radial.plot(KM$centers[1,],
            labels=names(KM$centers[1,]),
            rp.type="s",
            radial.lim=c(0,8),
            point.symbols=13,
            point.col="red",
            mar = c(2,1,5,2))


plot(ST_KM %>% select(age, absences), col = KM$cluster)
points(as.data.frame(KM$centers) %>% select(age, absences), col = 1:3, pch = 8, cex = 2)


ST_FIN <- ST_MOD %>% mutate(cluster_id = KM$cluster)
kable(head(ST_FIN))
```

Aunque es posible que haber normalizado los valores de las variables entre 1:5 puede que nos haya limitado la vision grafica de las dispersiones, en general vemos, tras un proceso iterativo en la busqueda de centros. Todo apunta a que aquellas variables como G1,G2,“studytime”, “traveltime” se encuentran en el grueso de valores máximos asociados a notas altas –> Y las variables de tipo Walc,Dalc,“goout” Tienen mayor influencia en el sentido inverso, y están asociados a la caida de notas. Mis fuentes principales son el ‘table(ST_MOD\(G3, KM\)cluster)’ y el esquema de centros radial.

6. APLICACION DE ALGORITMOS PREDICTIVOS

1. ARBOL DE DECISIONES

Usaremos el árbol de decisiones para este análisis predictivo debido a la simplicidad de la presentación de datos en nuestro Dataset, que nos permite hacer una rápida lectura y comparativa en torno a los resultados que arroje el algoritmo tras aplicarse.

La mexcla de variables categóricas y numéricas hace más complicado usar la metodología de regresiones lineales, lo que nos lleva a elegir este método como el principal.

Se utilizará el paquete Rpart, que es con el que hemos aprendido en la lección 3 y que parece encajar perfectamente con nuestro caso, utilizando todas las variables que hemos usado en el analisis

```
library(caret)
ST_new<- ST%>%select(sex, age, address,Pstatus, Medu, Fedu, Mjob, Fjob,studytime,traveltime,failures,higher,
internet,
                        goout, Dalc,Walc,health, absences,G1, G2, G3)
ARBOL <- rpart(G3 ~ .,
               data = ST_new,
               method = "class")
PRIN <- varImp(ARBOL)
rownames(PRIN)[order(PRIN$Overall, decreasing=TRUE)]

printcp(ARBOL)


plotcp(ARBOL)
```

Encontramos que G1 y G2 examen son predictores clave seguidos por niveles de asistencia, consumo de alcohol y trabajos de los padres.

La lógica del árbol consiste en sólo utilizar “attendance, Fjob, G1 y G2” como variables basadas en la correlación y la colinealidad entre algunas de las otras variables. –>

7. CONCLUSIONES

El hecho de que muchas variables estén correlacionadas es directamente proporcional a la baja significancia o capacidad en las predicciones finales, ya que no consiguen destacarse como variables unicas.

Respecto al análisis exploratorio según tecnicas de visualización o cruce de datos y aplicando clustering, se obtienen algunos resultados contradictorios, o quizás podría decirse que confirmatorios, como ocurre con el caso de la relación de ausencias, Alcohol entre y en fin de semana, o salidas,

Así como variables que influyen menos en las notas finales y tienen una repercusión escasa

Aun queda mucho por aprender en este fascinante campo. Con más tiempo y exhaustividad modelos supervisados y no supervisados podrían converger con satisfacción y mejorar la calidad de la certidumbre de los datos