

# Personas\_Dag

Josue Tapia

```
set.seed(12345)
data=read.csv("data_persona.csv")
data <- as.data.frame(data)
tsdem=read.csv("tsdem_eod2017/conjunto_de_datos/tsdem.csv")
head(data)
```

	id_soc	dias	sexo	edad	estrato	entidad	used_auto	used_autobusM1	used_bicicleta
1	1268	2	2	26	3	9	0	0	0
2	1269	2	2	22	3	9	0	0	0
3	1270	2	1	64	3	9	1	0	0
4	1271	1	2	41	3	9	0	0	0
5	1272	2	1	13	3	9	1	0	0
6	1273	2	2	56	3	9	0	0	0

	used_autobus	used_caminar	veces_auto
1	0	1	0
2	0	1	0
3	0	1	4
4	0	1	0
5	0	1	2
6	0	1	0

```
data=merge(data, tsdem, by = "id_soc")

data = data[, !names(data) %in% c("edad.x", "id_soc", "dias","id_hog","n_ren","parentesco","")]
#Borramos columnas inutilles
# No borrar P_3_7(ocupacion),tloc

data$veces_auto[data$veces_auto >2 ] = "True" #Si usa mas de dos veces el auto al dia True
data$veces_auto[data$veces_auto <=2 ] = "False" #DE otra manera Falso
data$veces_auto = as.factor(data$veces_auto)
```

```

names(data)[names(data) == "p3_7"] = "Ocupacion"
names(data)[names(data) == "sexo.x"]="sexo"
names(data)[names(data) == "estrato.x"]="estrato"
#Cambiar nombre de columnas

data$sexo= factor(data$sexo, levels= c(1,2), labels= c ("Hombre","Mujer"))

data$estrato = factor(data$estrato, levels = c(1,2,3,4), labels = c("Bajo","Medio_bajo","Med.
data$entidad[data$entidad != 9 & data$entidad != 15] = 0
data$entidad = factor(data$entidad,
                      levels = c(9, 15,0),
                      labels = c("Mexico", "Ciudad de Mexico","Other"))

data$used_auto= factor(data$used_auto , levels = c(0,1), labels = c("False","True"))
data$used_autobusM1= factor(data$used_autobusM1 , levels = c(0,1), labels = c("False","True"))
data$used_autobus= factor(data$used_autobus , levels = c(0,1), labels = c("False","True"))
data$used_bicicleta= factor(data$used_bicicleta , levels = c(0,1), labels = c("False","True"))
data$used_caminar = factor(data$used_caminar, levels = c(0,1), labels = c("False","True"))
data$Ocupacion[!data$Ocupacion %in% 1:8] = 0
data$Ocupacion= factor(data$Ocupacion, levels = c(1,2,3,4,5,6,7,8,0), labels = c(
  "trabajó",
  "tenía trabajo, pero no trabajó",
  "buscó trabajo",
  "Es estudiante",
  "Se dedica a los quehaceres del hogar o a cuidar a sus hijos",
  "Es jubilado(a) o pensionado(a)",
  "Está incapacitado(a) permanentemente para trabajar",
  "No trabajó",
  "Other"
))

data$tloc = factor(data$tloc , levels = c(1,2,3,4),labels = c ("Localidades mayores de 10000
"Localidades de 15,000 a 99,999 habitantes",
"Localidades de 2,500 a 14,999 habitantes",
"Localidades menores de 2,500 habitantes"
))

#Convertir labels a lenguaje hablado

```

```
head(data)
```

	sexo	estrato	entidad	used_auto	used_autobusM1	used_bicicleta
1	Hombre	Medio_bajo	Mexico	False	False	False
2	Mujer	Medio_bajo	Mexico	False	False	False
3	Mujer	Medio_bajo	Mexico	False	False	False
4	Mujer	Medio_bajo	Mexico	False	False	False
5	Hombre	Medio_bajo	Mexico	False	False	False
6	Mujer	Medio_bajo	Mexico	False	False	False

	used_autobus	used_caminar	veces_auto
1	False	True	False
2	False	True	False
3	False	True	False
4	False	True	False
5	False	True	False
6	False	True	False

	Ocupacion
1	trabajó
2	trabajó
3	Other
4	No trabajó
5	trabajó
6	Se dedica a los quehaceres del hogar o a cuidar a sus hijos

	tloc
1	Localidades mayores de 100000 habitantes
2	Localidades mayores de 100000 habitantes
3	Localidades mayores de 100000 habitantes
4	Localidades mayores de 100000 habitantes
5	Localidades mayores de 100000 habitantes
6	Localidades mayores de 100000 habitantes

```
write.csv(data, "data_persona_limpia.csv", row.names = FALSE)
```

## DAG 1

```
library(bnlearn)
library(Rgraphviz)
```

Loading required package: graph

Loading required package: BiocGenerics

Loading required package: generics

Attaching package: 'generics'

The following objects are masked from 'package:base':

as.difftime, as.factor, as.ordered, intersect, is.element, setdiff,  
setequal, union

Attaching package: 'BiocGenerics'

The following object is masked from 'package:bnlearn':

score

The following objects are masked from 'package:stats':

IQR, mad, sd, var, xtabs

The following objects are masked from 'package:base':

anyDuplicated, aperm, append, as.data.frame, basename, cbind,  
colnames, dirname, do.call, duplicated, eval, evalq, Filter, Find,  
get, grep, grepl, is.unsorted, lapply, Map, mapply, match, mget,  
order, paste, pmax, pmax.int, pmin, pmin.int, Position, rank,  
rbind, Reduce, rownames, sapply, saveRDS, table, tapply, unique,  
unsplit, which.max, which.min

Attaching package: 'graph'

The following objects are masked from 'package:bnlearn':

degree, nodes, nodes<-

Loading required package: grid

```

# Definir DAG
nodes <- c("sexo","estrato","entidad","tloc",
           "Ocupacion","used_auto","used_autobusM1",
           "used_autobus","used_bicicleta","used_caminar","veces_auto")

# Grafo vacío
dag <- empty.graph(nodes)

# Definir arcos directamente como matriz
arcs(dag) <- matrix(c(
  # Ocupacion depende de sexo, estrato, entidad, tloc
  "sexo","Ocupacion",
  "estrato","Ocupacion",
  "entidad","Ocupacion",
  "tloc","Ocupacion",

  # Medios de transporte dependen de sexo, estrato, entidad, tloc, Ocupacion
  "sexo","used_auto",
  "estrato","used_auto",
  "entidad","used_auto",
  "tloc","used_auto",
  "Ocupacion","used_auto",

  "sexo","used_autobusM1",
  "estrato","used_autobusM1",
  "entidad","used_autobusM1",
  "tloc","used_autobusM1",
  "Ocupacion","used_autobusM1",

  "sexo","used_autobus",
  "estrato","used_autobus",
  "entidad","used_autobus",
  "tloc","used_autobus",
  "Ocupacion","used_autobus",

  "sexo","used_bicicleta",
  "estrato","used_bicicleta",
  "entidad","used_bicicleta",
  "tloc","used_bicicleta",
  "Ocupacion","used_bicicleta",

  "sexo","used_caminar",

```

```

"estrato","used_caminar",
"entidad","used_caminar",
"tloc","used_caminar",
"Ocupacion","used_caminar",

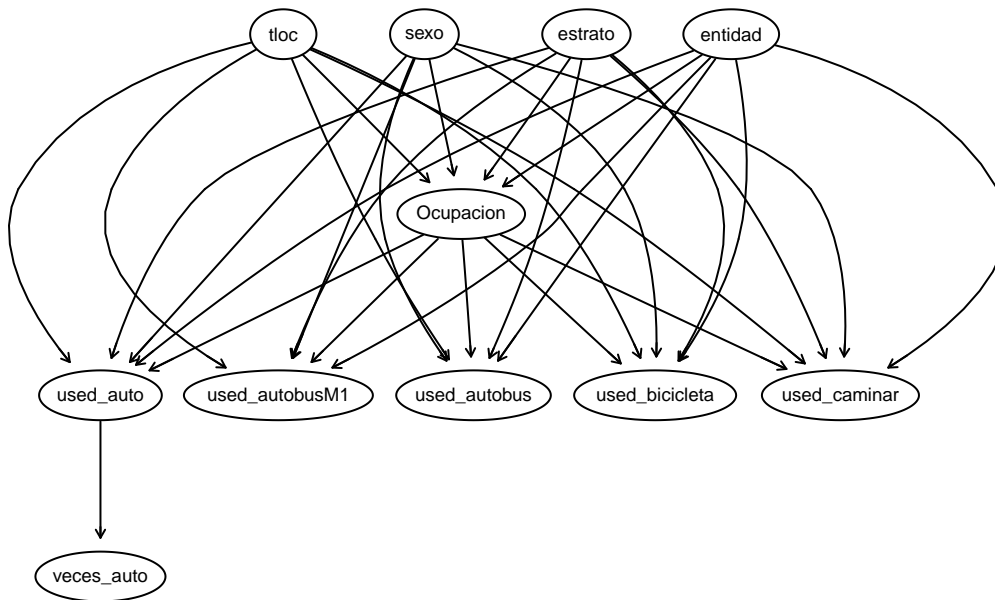
"used_auto","veces_auto"
), ncol = 2, byrow = TRUE,
dimnames = list(NULL, c("from","to")))

```

```

graphviz.plot(dag, shape = "ellipse") #Referencias : https://www.tandfonline.com/doi/full/10

```



```

#Articulo sobre ocupacion

```

```

#Referencias : https://repository.eafit.edu.co/server/api/core/bitstreams/710b4da3-8e42-43cf

```

```

#Las referencias hablan de como los factores sociales determinan la ocupacion y esta misma m

```

```

score(dag,data= data , type = "bic") #Score de la primer DAG

```

```

[1] -976422.6

```

```

#DAG 2

```

```

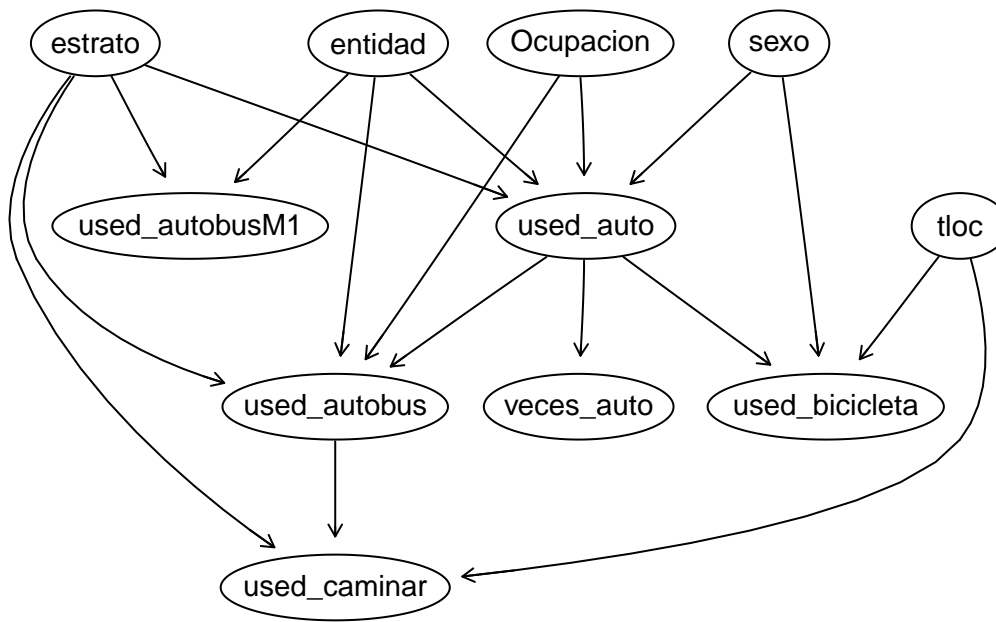
nodes <- c("sexo","estrato","entidad","tloc",
           "Ocupacion","used_auto","used_autobusM1",
           "used_autobus","used_bicicleta","used_caminar","veces_auto")

dag <- empty.graph(nodes)

arcs(dag) <- matrix(c(
  "sexo","used_auto",
  "sexo","used_bicicleta",
  "estrato","used_auto",
  "estrato","used_autobusM1",
  "estrato","used_autobus",
  "estrato","used_caminar",
  "entidad","used_auto",
  "entidad","used_autobusM1",
  "entidad","used_autobus",
  "tloc","used_bicicleta",
  "tloc","used_caminar",
  "Ocupacion","used_auto",
  "Ocupacion","used_autobus",
  "used_auto","used_autobus",
  "used_auto","used_bicicleta",
  "used_autobus","used_caminar",
  "used_auto","veces_auto"
), ncol = 2, byrow = TRUE,
dimnames = list(NULL, c("from","to")))

graphviz.plot(dag, layout = "dot", shape = "ellipse") #Referencias: Está basado en lo mismo d

```



```
score(dag,data= data , type = "bic")
```

```
[1] -972143.3
```

Por lo tanto la DAG 1 es mejor que la DAG2 esto puede ser por el hecho de que la ocupación si es dependiente de los factores sociales jsuto como el artículo lo describe

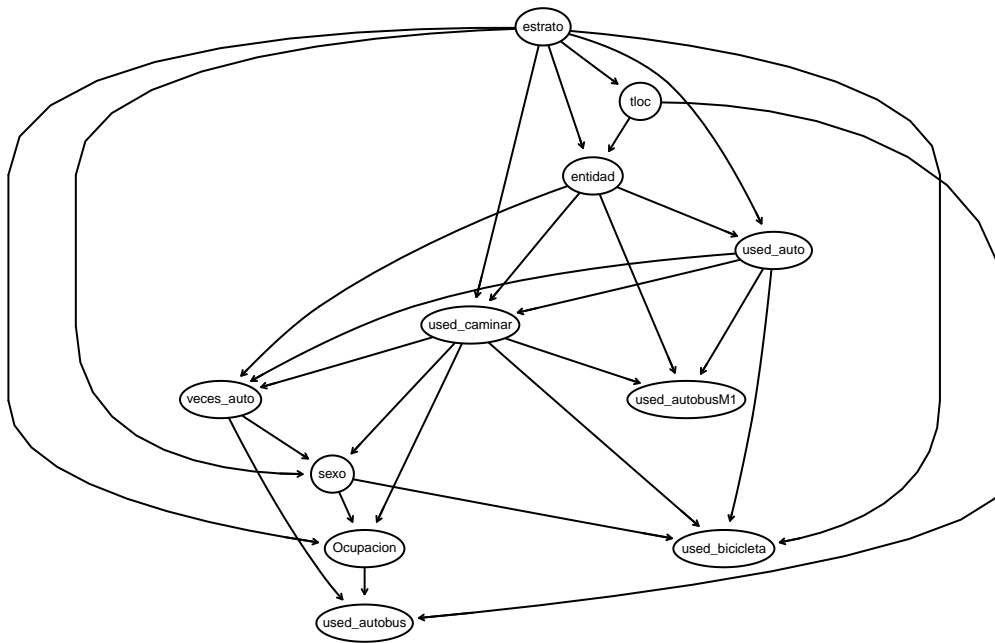
## DAG3

```
best_dag= hc(data,type = "aic") #Obtenemos la mejor dag con hilll climbing
```

```
Warning in check.unused.args(extra, c(learning.extra.args[[heuristic]], :
unused argument(s): 'type'.
```

```
# Dibujar
graphviz.plot(best_dag, shape="ellipse")
```





```
score(best_dag,data= data , type = "bic") #Score de la mejor DAG
```

```
[1] -913248
```

```
#Fittear la mejor DAG
fitted_best_dag=bn.fit(best_dag,data = data)
```

```
#Query 1
```

**¿Cuál es la probabilidad de que una persona del estrato socioeconómico más bajo camine o use bicicleta como medio de transporte versus una persona del estrato más alto?**

```
p1=cpquery(fitted_best_dag,event =(used_bicicleta=="True" | used_caminar=="True"), evidence =
p2=cpquery(fitted_best_dag,event =(used_bicicleta=="True" | used_caminar=="True"), evidence =
cat("Probabilidad de que alguien de estrato bajo camine o use bicicleta: ", p1, "\n")
```

```
Probabilidad de que alguien de estrato bajo camine o use bicicleta: 0.8484848
```

```
cat("Probabilidad de que alguien de estrato alto camine o use bicicleta: ", p2, "\n")
```

Probabilidad de que alguien de estrato alto camine o use bicicleta: 0.4672464

###Por lo tanto es mucho más probable que use bicicleta o camine alguien del estrato más bajo que de estrato más alto

#Query 2

**¿Cuál es la probabilidad de que una persona de sexo femenino de la entidad de Mexico que habita en una localidad de 15,000 a 99,999 habitantes utilice un autobús?**

```
cpquery(fitted_best_dag, event=(used_autobus== "True" | used_autobusM1== "True"),evidence =
```

```
[1] 0.1157025
```

###Esto nos indica que sooo hay una probabilidad del 11% de personas las cuales tomaran el camión y tiene sentido porque vienen de una localidad muy pequeña y ahí casi no hay transporte público

#Query 3

**¿Cuál es la probabilidad de que una estudiante mujer haga más de 2 viajes al día en un automóvil?**

```
p1=cpquery(fitted_best_dag, event=(veces_auto == "True"),evidence = (Ocupacion == "Es estud.
```

```
cat("Probabilidad de que una estudiante mujer haga más de 2 viajes al día en un automóvil: "
```

Probabilidad de que una estudiante mujer haga más de 2 viajes al día en un automóvil: 0.0702

###Esta probabilidad tan baja en parte es debido a que la encuesta tiene muchos mas observaciones del estrato bajo que del alto esto provoca que tales personas incluso no poseen un auto propio. Aunque si lo cambiamos a que sea una mujer estudiante de estrato alto tendremos:

```
p=cpquery(fitted_best_dag, event=(veces_auto == "True") ,evidence = (Ocupacion == "Es estudiante"))  
  
cat("Probabilidad de que una estudiante mujer de estrato alto haga más de 2 viajes al día en un auto propio")
```

Probabilidad de que una estudiante mujer de estrato alto haga más de 2 viajes al día en un auto propio

**Aumenta bastante la probabilidad**