TRABAJO DE GRADO
Opción Seminario-Diplomado.

Proyectos Cursos Crehana de Machine Learning e Inteligencia Artificial

Corporación Universitaria Remington.
Facultad de ingenierías.
Ingeniería de sistemas.

Judy Katerine Ríos Ortiz.
Tutor: Juan Pablo Vélez Uribe
Opción de Trabajo de grado Seminario-Diplomado.
2024

Tabla de Contenidos

Proyectos Cursos Crehana de Machine Learning e Inteligencia Artificial

Proyecto final de Introducción a Machine Learning

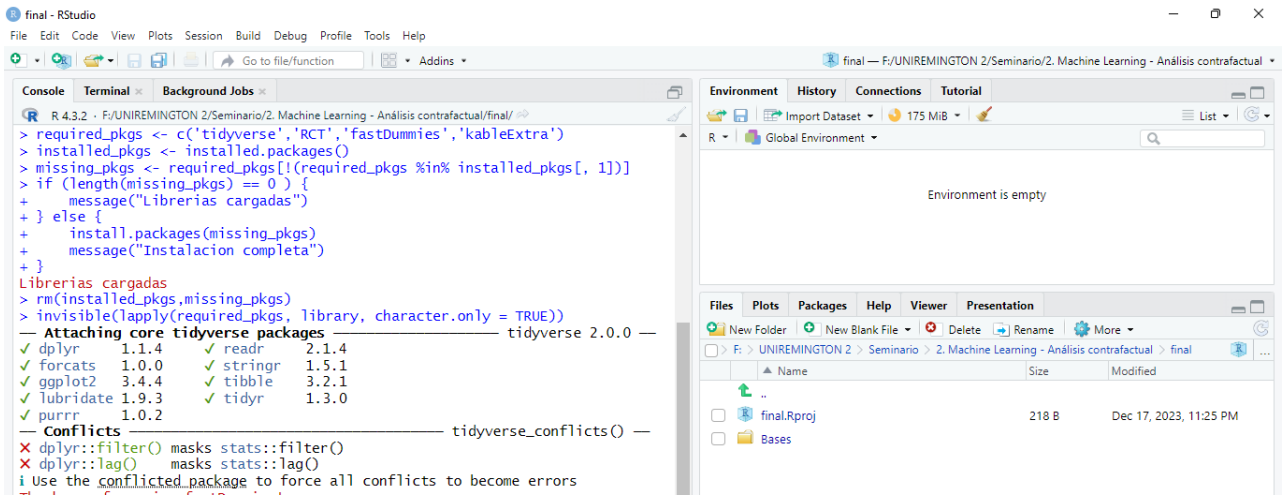
Proyecto final de Machine Learning: Análisis contrafactual

Avance 1

El objetivo de este avance es realizar una asignación aleatoria estratificada de los clientes inactivos. Esto con la finalidad de determinar los grupos de clientes que recibirán un tipo de comunicación específica en la estrategia experimental.

Instalación de librerías

```
required_pkgs <- c('tidyverse','RCT','fastDummies','kableExtra')
installed_pkgs <- installed.packages()
missing_pkgs <- required_pkgs[!(required_pkgs %in% installed_pkgs[, 1])]
if (length(missing_pkgs) == 0 ) {
message("Librerias cargadas")
} else {
install.packages(missing_pkgs)
message("Instalacion completa")
}
rm(installed_pkgs,missing_pkgs)
invisible(lapply(required_pkgs, library, character.only = TRUE))
rm(required_pkgs)
```



Pregunta 1

Explora la base de datos de clientes nuevos. ¿A qué nivel de desagregación está la base? ¿Cuántos clientes únicos? ¿Qué variables tienen valores vacíos? Decide si debes excluir a esas observaciones o mantenerlas y justifica tu decisión.

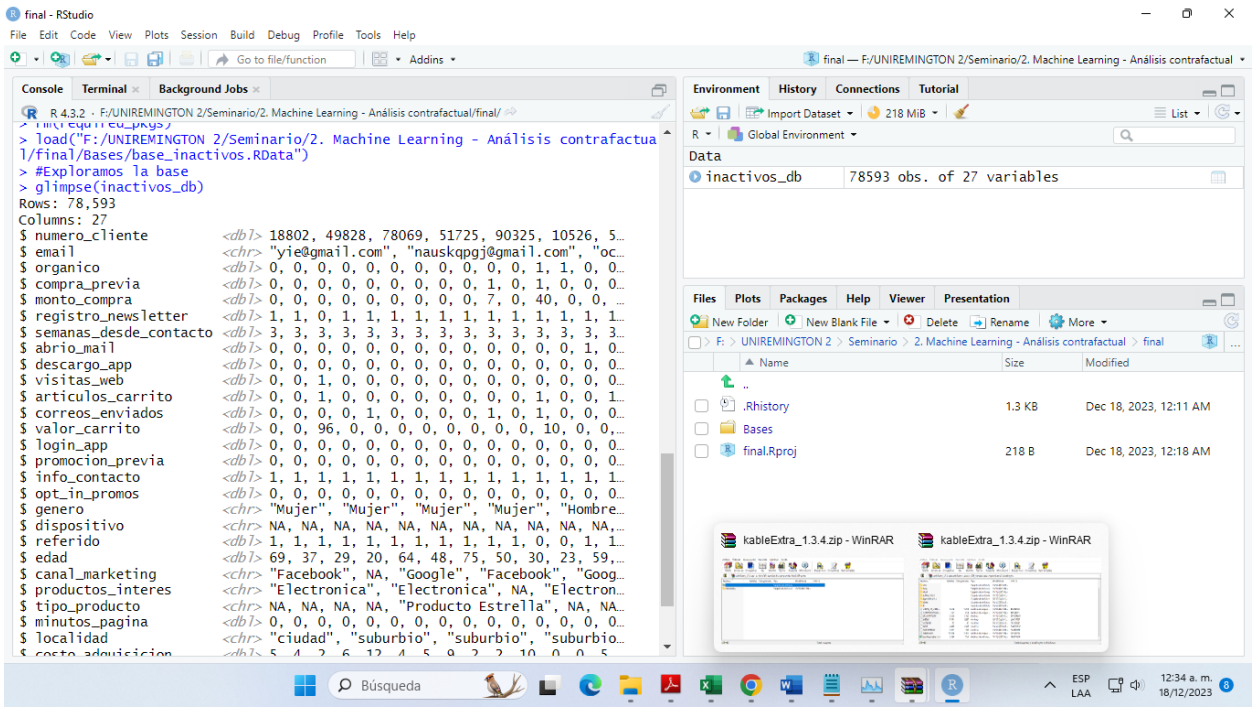
Respuesta: Aunque las variables como el tipo de dispositivo, canal de marketing, productos de interés y tipo de producto puedan tener valores nulos, la carencia de información no obstaculiza el procedimiento de asignación.

#Cargamos la base

load("F:/UNIREMINGTON 2/Seminario/2. Machine Learning - Análisis contrafactual/final/Bases/base_inactivos.RData")

#Exploramos la base

glimpse(inactivos_db)



 # La base está a nivel cliente (numero_cliente). Con 78,593 observaciones únicas.

length(unique(inactivos_db$numero_cliente))

```
> # La base está a nivel cliente (numero_cliente). Con 78,593 observaciones úni
cas.
> length(unique(inactivos_db$numero_cliente))
[1] 78593
```

 # Valores faltantes

missings<-map_dbl(inactivos_db %>% select_all(),

~100*sum(is.na(.))/nrow(inactivos_db))

missings[missings>0]

```
> # Valores faltantes
> missings<-map_dbl(inactivos_db %>% select_all(),
+                   ~100*sum(is.na(.))/nrow(inactivos_db))
> missings[missings>0]
      dispositivo   canal_marketing productos_interes     tipo_producto
          98.54694          52.96019          16.78394          81.97677
```

Pregunta 2:

¿Qué variables crees que puedan estar más correlacionadas con el impacto del tratamiento?

Respuesta:
Las variables como orgánico, registro_newsletter, abrio_mail, descargo_app y visitas web ofrecen indicios sobre el interés del usuario en adquirir un producto, lo que sugiere su relevancia en la estratificación para eliminar posibles sesgos. Además, variables como el costo de adquisición y la presencia de productos en el carrito de compra son relevantes para la evaluación de impacto. Estratificar por costo de adquisición permitirá determinar la rentabilidad de usuarios con mayores costos. La variable de productos en el carrito proporcionará insights sobre si la promoción fue un incentivo decisivo para completar la compra.

Pregunta 3:
Realiza una asignación aleatoria de la población de clientes en 3 grupos de tamaño similar. El grupo asignadodeterminará el tipo de promoción que recibirán.

```
#Variables estratificadoras
inactivos_db <- inactivos_db %>%
mutate(grupo_cac = ntile(costo_adquisicion, 4),
agrego_articulo = if_else(articulos_carrito==0,0,1))

# Asignación
asignacion <- treatment_assign(inactivos_db,
share_control = .33,
n_t=2,
strata_varlist = vars(organico, registro_newsletter,
abrio_mail, descargo_app,
```

```
agrego_articulo, grupo_cac),
seed = 2000,
key = 'numero_cliente')
list2env(asignacion, envir = .GlobalEnv)
## <environment: R_GlobalEnv>
# Juntamos la base de clientes inactivos con el universo asignado.
inactivos_db <- left_join(inactivos_db,
data,
by = "numero_cliente")
```

Pregunta 4.

Realiza las pruebas de balance sobre todas las variables. ¿Están balanceadas las variables entre los 3 grupos?

```
# Transformar las variables de texto en categóricas para poder incluirlas
# en la prueba de balance
inactivos_db_aux <- inactivos_db %>%
dummy_cols(select_columns = c("genero","dispositivo","canal_marketing",
"productos_interes","tipo_producto","localidad"),
ignore_na = T, remove_selected_columns = T) %>%
mutate_at(vars(c(starts_with(c("genero","dispositivo","canal_marketing",
"productos_interes","tipo_producto","localidad")))),
function(x) x = if_else(is.na(x),0,as.double(x)))

# Tabla de balance
balance_tab <- balance_table(data = inactivos_db_aux %>%
select(-c(numero_cliente, email)),
treatment = 'treat')
kable(balance_tab,
caption = "Tabla de balance",
digits = 2)
```

Guardamos la base en un archivo csv para poder compartirlo con el equipo de marketing y programar los
envios. El arhivo especifica el grupo de pertenencia del usuario y el tratamiento a recibir.

```
inactivos_marketing <- inactivos_db%>%
select(numero_cliente,email,treat)%>%
mutate(treat = case_when(treat==0 ~ "Sin Comunicacion",
treat==1 ~ "Promocion Cashback",
treat==2 ~ "Promocion Descuento"))
#fwrite(inactivos_marketing, file = "Bases output/inactivos_marketing.csv")
```

Table 1: Tabla de balance

variables1	Media_control1	Media_trat1	Media_trat2	p_value1	p_value2
abrio_mail	0.15	0.15	0.15	0.84	0.72
agrego_articulo	0.20	0.20	0.21	0.78	0.70
articulos_carrito	0.39	0.39	0.40	0.82	0.36
canal_marketing_Facebook	0.35	0.35	0.35	0.37	0.60
canal_marketing_Google	0.08	0.08	0.08	0.95	0.90
canal_marketing_Instagram	0.04	0.03	0.04	0.07	0.57
canal_marketing_Otro	0.01	0.01	0.00	0.55	0.05
compra_previa	0.12	0.12	0.12	0.05	0.03
correos_enviados	0.24	0.24	0.24	0.28	0.76
costo_adquisicion	3.54	3.55	3.54	0.69	0.97
descargo_app	0.01	0.01	0.02	0.62	0.30
dispositivo_Android	0.00	0.00	0.00	0.05	0.17
dispositivo_IOS	0.01	0.01	0.01	0.57	0.69
edad	50.40	50.44	50.50	0.85	0.55
genero_Hombre	0.36	0.35	0.36	0.73	0.83
genero_Mujer	0.64	0.65	0.64	0.73	0.83
grupo_cac	2.50	2.50	2.50	0.98	0.85
info_contacto	0.56	0.57	0.56	0.14	0.42
localidad_ciudad	0.45	0.45	0.45	0.63	0.74
localidad_comunidad	0.34	0.34	0.34	0.41	0.83
localidad_extranjero	0.01	0.01	0.01	0.19	0.75
localidad_suburbio	0.20	0.20	0.20	0.23	0.55
login_app	0.11	0.10	0.11	0.63	0.97
minutos_pagina	0.44	0.45	0.45	0.46	0.20
missfit	0.06	0.06	0.06	0.99	0.99
monto_compra	25.83	25.94	26.52	0.92	0.58
opt_in_promos	0.33	0.33	0.33	0.17	0.62
organico	0.49	0.49	0.49	0.98	0.73
productos_interes_Deportes	0.03	0.03	0.02	0.22	0.02
productos_interes_Electronica	0.59	0.59	0.59	0.62	0.62
productos_interes_Gaming	0.19	0.19	0.19	0.82	0.99
productos_interes_Hogar	0.03	0.03	0.03	0.35	0.33
promocion_previa	0.00	0.00	0.00	0.67	0.82
referido	0.51	0.51	0.51	0.99	0.81
registro_newsletter	0.83	0.83	0.83	0.97	0.94
semanas_desde_contacto	2.24	2.23	2.23	0.41	0.28
strata	27.45	27.45	27.50	0.98	0.74
tipo_producto_Producto Estrella	0.18	0.17	0.17	0.22	0.14
tipo_producto_Producto Nuevo	0.00	0.00	0.00	0.22	0.55
tipo_producto_Producto Temporada	0.00	0.00	0.00	0.40	0.78
valor_carrito	10.43	10.56	10.81	0.74	0.33
visitas_web	0.55	0.56	0.56	0.54	0.73

Avance 2

El objetivo de este avance es realizar una evaluación de impacto del experimento y determinar qué opción fue más efectiva para incrementar las ventas y para qué perfil de cliente.

Instalación de librerías

```
required_pkgs <- c('tidyverse','RCT','writexl',"scales")
installed_pkgs <- installed.packages()
missing_pkgs <- required_pkgs[!(required_pkgs %in% installed_pkgs[, 1])]
if (length(missing_pkgs) == 0 ) {
message("Librerias cargadas")
} else {
install.packages(missing_pkgs)
message("Instalacion completa")
}
rm(installed_pkgs,missing_pkgs)
invisible(lapply(required_pkgs, library, character.only = TRUE))
rm(required_pkgs)
```

Pregunta 1

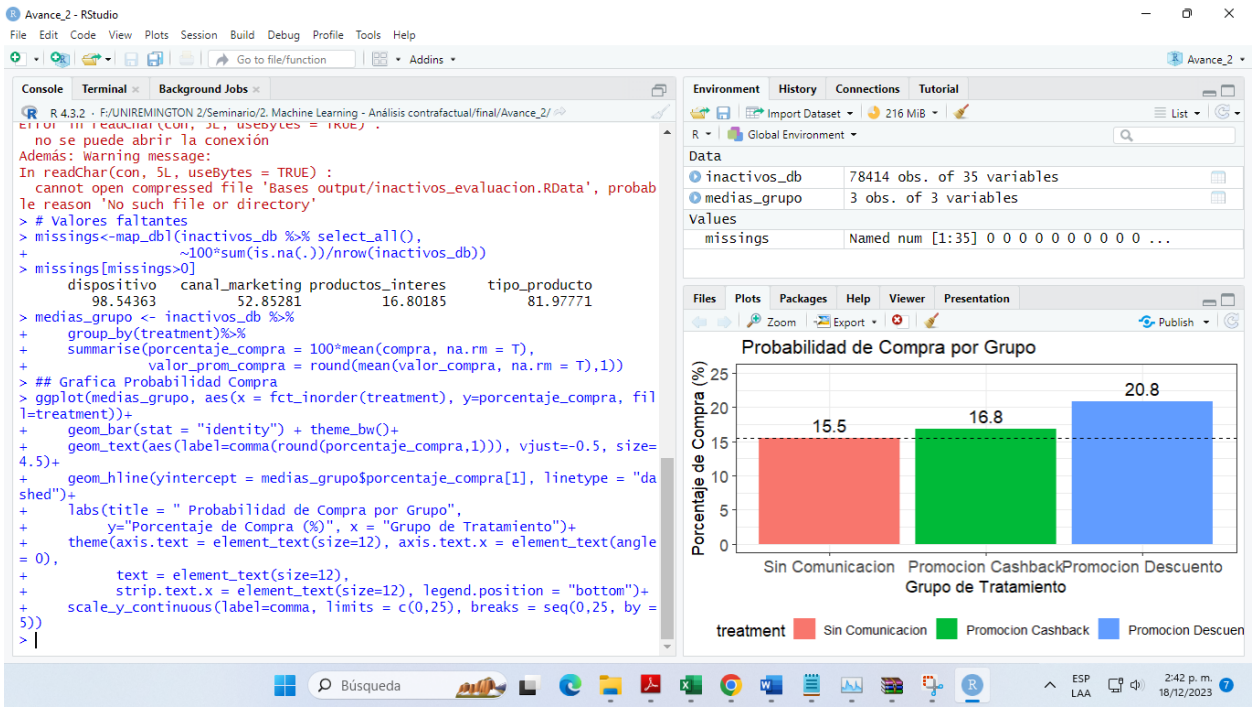
Realiza una comparación del porcentaje de conversiones y el valor promedio de las ventas entre los 3 distintos grupos de clientes. ¿Observas alguna diferencia entre los grupos?

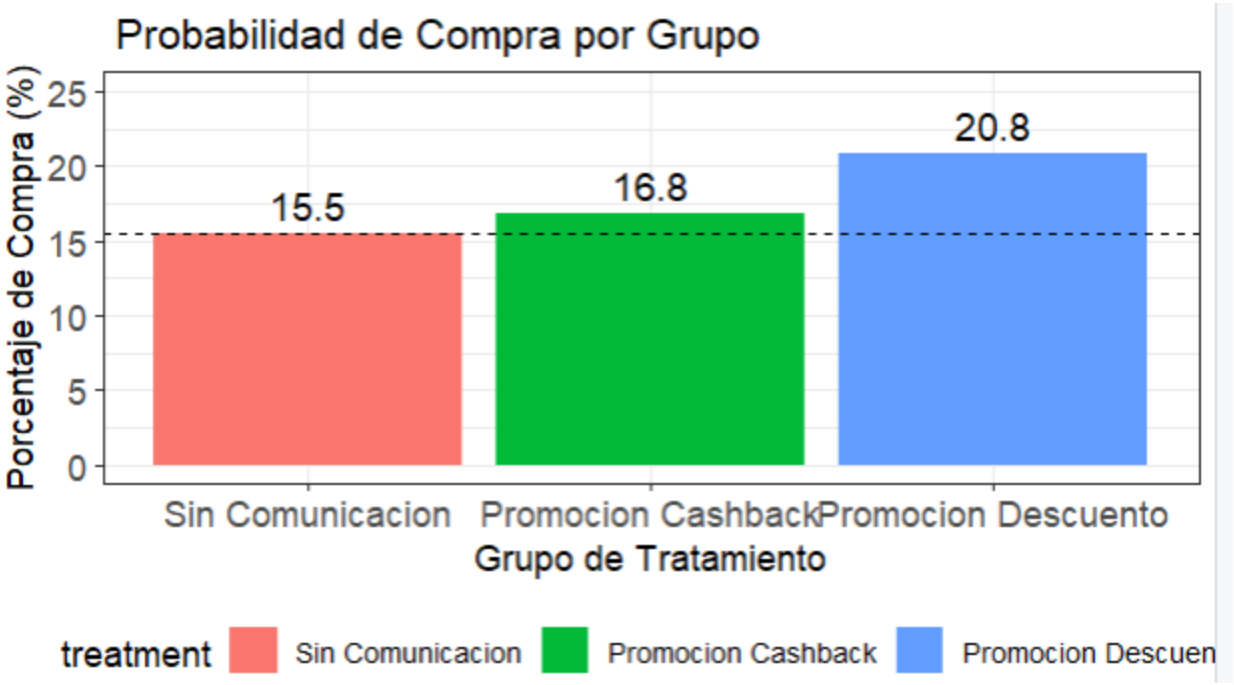
Respuesta: Se observa que el grupo que recibió la promoción del descuento exhibió un porcentaje más alto de conversiones y también montos de compra más elevados.

```
#Cargamos la base
load("Bases output/inactivos_evaluacion.RData")

# Valores faltantes
missings<-map_dbl(inactivos_db %>% select_all(),
~100*sum(is.na(.))/nrow(inactivos_db))
missings[missings>0]
```

```
> # Valores faltantes
> missings<-map_dbl(inactivos_db %>% select_all(),
+                   ~100*sum(is.na(.))/nrow(inactivos_db))
> missings[missings>0]
      dispositivo   canal_marketing productos_interes     tipo_producto
         98.54363          52.85281          16.80185          81.97771
```

```
#Generamos la comparativa de medias
medias_grupo <- inactivos_db %>%
group_by(treatment)%>%
summarise(porcentaje_compra = 100*mean(compra, na.rm = T),
valor_prom_compra = round(mean(valor_compra, na.rm = T),1))

## Grafica Probabilidad Compra
ggplot(medias_grupo, aes(x = fct_inorder(treatment), y=porcentaje_compra,
fill=treatment))+
geom_bar(stat = "identity") + theme_bw()+
geom_text(aes(label=comma(round(porcentaje_compra,1))), vjust=-0.5, size= 4.5)+
geom_hline(yintercept = medias_grupo$porcentaje_compra[1], linetype =
"dashed")+
labs(title = " Probabilidad de Compra por Grupo",
y="Porcentaje de Compra (%)", x = "Grupo de Tratamiento")+
theme(axis.text = element_text(size=12), axis.text.x = element_text(angle = 0),
text = element_text(size=12),
strip.text.x = element_text(size=12), legend.position = "bottom")+
scale_y_continuous(label=comma, limits = c(0,25), breaks = seq(0,25, by =5))
```


Avance_2 - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Console Terminal Background Jobs
R 4.3.2 · F:/UNIREMINGTON 2/Seminario/2. Machine Learning - Análisis contrafactual/final/Avance_2/
no se puede abrir la conexión
Además: Warning message:
In readChar(con, 5L, useBytes = TRUE) :
cannot open compressed file 'Bases output/inactivos_evaluacion.RData', probable reason 'No such file or directory'
> # Valores faltantes
> missings<-map_dbl(inactivos_db %>% select_all(),
+ ~100*sum(is.na(.))/nrow(inactivos_db))
> missings[missings>0]
dispositivo canal_marketing productos_interes tipo_producto
98.54363 52.85281 16.80185 81.97771
> medias_grupo <- inactivos_db %>%
+ group_by(treatment)%>%
+ summarise(porcentaje_compra = 100*mean(compra, na.rm = T),
+ valor_prom_compra = round(mean(valor_compra, na.rm = T),1))
> ## Grafica Probabilidad Compra
> ggplot(medias_grupo, aes(x = fct_inorder(treatment), y=porcentaje_compra, fill=treatment))+
+ geom_bar(stat = "identity") + theme_bw()+
+ geom_text(aes(label=comma(round(porcentaje_compra,1))), vjust=-0.5, size=4.5)+
+ geom_hline(yintercept = medias_grupo$porcentaje_compra[1], linetype = "dashed")+
+ labs(title = " Probabilidad de Compra por Grupo",
+ y="Porcentaje de Compra (%)", x = "Grupo de Tratamiento")+
+ theme(axis.text = element_text(size=12), axis.text.x = element_text(angle = 0),
+ text = element_text(size=12),
+ strip.text.x = element_text(size=12), legend.position = "bottom")+
+ scale_y_continuous(label=comma, limits = c(0,25), breaks = seq(0,25, by = 5))
Environment History Connections Tutorial
Import Dataset 216 MiB
Global Environment
Data
inactivos_db 78414 obs. of 35 variables
medias_grupo 3 obs. of 3 variables
Values
missings Named num [1:35] 0 0 0 0 0 0 0 0 0 0 ...
Files Plots Packages Help Viewer Presentation
Zoom Export Publish
Probabilidad de Compra por Grupo
Porcentaje de Compra (%)
15.5
16.8
20.8
Sin Comunicacion Promocion Cashback Promocion Descuento
Grupo de Tratamiento
treatment Sin Comunicacion Promocion Cashback Promocion Descuen
Búsqueda
ESP LAA
2:42 p. m. 18/12/2023


Probabilidad de Compra por Grupo
Porcentaje de Compra (%)
25
20
15
10
5
0
15.5
16.8
20.8
Sin Comunicacion
Promocion Cashback
Promocion Descuento
Grupo de Tratamiento
treatment
Sin Comunicacion
Promocion Cashback
Promocion Descuen

Pregunta 2:

Estima una regresión de evaluación de impacto de los efectos de tratamiento (ITT). Incluye efectos fijos por estrato en tu especificación. Reporta en una tabla el efecto promedio para cada grupo de tratamiento y su significancia estadística correspondiente

```
resultados_itt <- impact_eval(data = inactivos_db,
endogenous_vars = c("compra", "valor_compra"),
treatment = "treat",
fixed_effect_vars = "strata")
list2env(resultados_itt, envir = .GlobalEnv)

## <environment: R_GlobalEnv>
write_xlsx(resultados_itt,"Bases output/resultados_itt.xlsx")

# Coeficientes de variable compra
coef_compra <- compra%>%
mutate(term = parse_number(term),
p.value_compra = str_c("p=", round(p.value, digits = 2)),
treatment = as.factor(if_else(term==1,"Promocion Cashback",
"Promocion Descuento")))%>%
rename(prob_compra = estimate) %>%
select(treatment, prob_compra, p.value_compra)

# Coeficientes de variable valor de compra
coef_valor_compra <- valor_compra%>%
mutate(term = parse_number(term),
p.value_valor_compra = str_c("p=", round(p.value, digits = 2)),
treatment = as.factor(if_else(term==1,"Promocion Cashback",
"Promocion Descuento")))%>%
rename(valor_compra = estimate) %>%
select(treatment, valor_compra, p.value_valor_compra)
```

```
impactos <- left_join(coef_compra, coef_valor_compra, by = "treatment")
medias_grupo <- left_join(medias_grupo, impactos, by = "treatment")
rm(list = setdiff(ls(), c("inactivos_db","medias_grupo")))
```

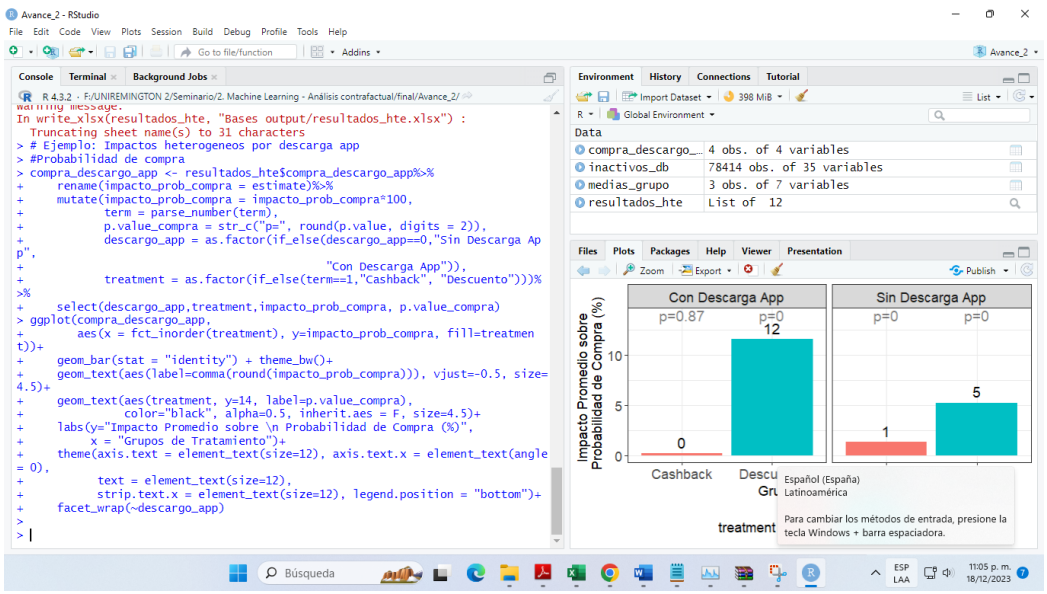
Pregunta 3.

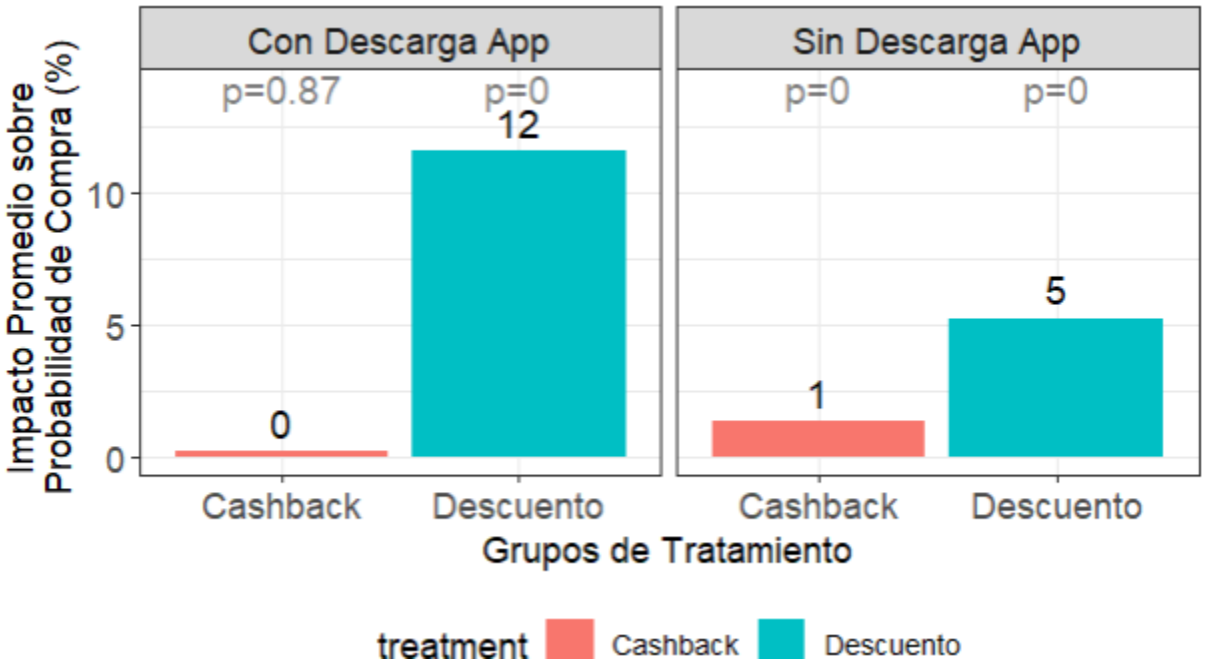
Realiza las pruebas de balance sobre todas las variables. ¿Están balanceadas las variables entre los 3 grupos?

Respuesta: Los clientes registrados al newsletter mostraron tasas de conversión más altas y montos de compra superiores en comparación con los no registrados. Además, los usuarios que descargaron la aplicación experimentaron un aumento del 11% en la probabilidad de compra y un valor de compra incrementado en 10 dólares en comparación con aquellos que no recibieron comunicación. En contraste, los clientes con un mayor costo de adquisición (CAC) exhibieron tasas de conversión más bajas y montos de compra inferiores en comparación con aquellos con un CAC menor.

```
resultados_hte <- impact_eval(data = inactivos_db,
endogenous_vars = c("compra", "valor_compra"),
treatment = "treat",
heterogenous_vars = c("organico", "registro_newsletter",
"descargo_app","agrego_articulo",g
"grupo_cac"),
fixed_effect_vars = "strata")
write_xlsx(resultados_hte,"Bases output/resultados_hte.xlsx")

# Ejemplo: Impactos heterogeneos por descarga app
#Probabilidad de compra
compra_descargo_app <- resultados_hte$compra_descargo_app%>%
rename(impacto_prob_compra = estimate)%>%
mutate(impacto_prob_compra = impacto_prob_compra*100,
term = parse_number(term),
```

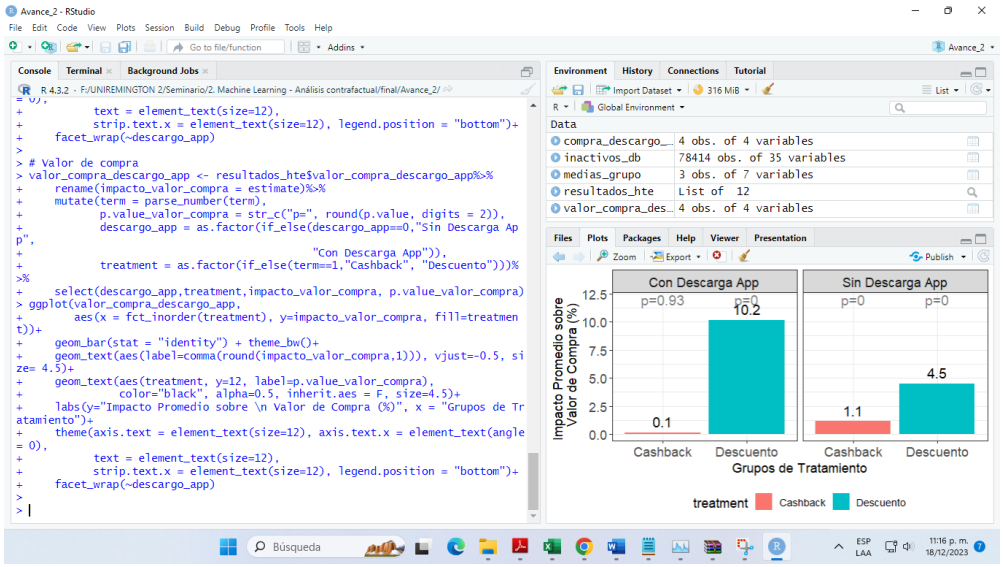
```
p.value_compra = str_c("p=", round(p.value, digits = 2)),
descargo_app = as.factor(if_else(descargo_app==0,"Sin Descarga App",
"Con Descarga App")),
treatment = as.factor(if_else(term==1,"Cashback", "Descuento")))%>%
select(descargo_app,treatment,impacto_prob_compra, p.value_compra)
ggplot(compra_descargo_app,
aes(x = fct_inorder(treatment), y=impacto_prob_compra, fill=treatment))+
geom_bar(stat = "identity") + theme_bw()+
geom_text(aes(label=comma(round(impacto_prob_compra))), vjust=-0.5, size= 4.5)+
geom_text(aes(treatment, y=14, label=p.value_compra),
color="black", alpha=0.5, inherit.aes = F, size=4.5)+
labs(y="Impacto Promedio sobre \n Probabilidad de Compra (%)",
x = "Grupos de Tratamiento")+
theme(axis.text = element_text(size=12), axis.text.x = element_text(angle = 0),
text = element_text(size=12),
strip.text.x = element_text(size=12), legend.position = "bottom")+
facet_wrap(~descargo_app)
```

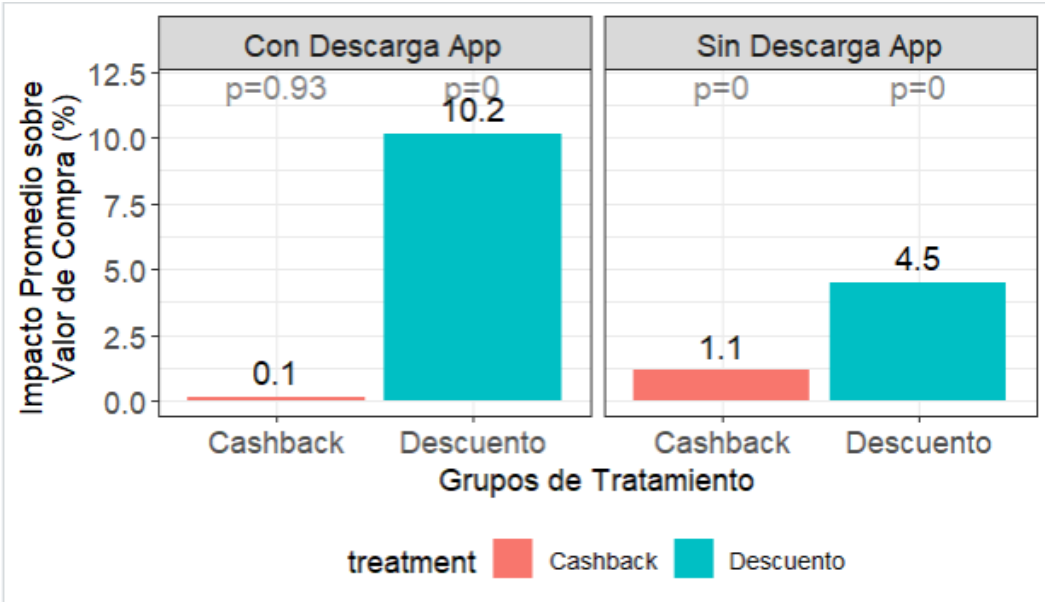





```
# Valor de compra
valor_compra_descargo_app <- resultados_hte$valor_compra_descargo_app%>%
rename(impacto_valor_compra = estimate)%>%
mutate(term = parse_number(term),
p.value_valor_compra = str_c("p=", round(p.value, digits = 2)),
descargo_app = as.factor(if_else(descargo_app==0,"Sin Descarga App",
"Con Descarga App")),
treatment = as.factor(if_else(term==1,"Cashback", "Descuento")))%>%
select(descargo_app,treatment,impacto_valor_compra, p.value_valor_compra)
ggplot(valor_compra_descargo_app,
aes(x = fct_inorder(treatment), y=impacto_valor_compra, fill=treatment))+
geom_bar(stat = "identity") + theme_bw()+
geom_text(aes(label=comma(round(impacto_valor_compra,1))), vjust=-0.5, size= 4.5)+
geom_text(aes(treatment, y=12, label=p.value_valor_compra),
color="black", alpha=0.5, inherit.aes = F, size=4.5)+
```

```
labs(y="Impacto Promedio sobre \n Valor de Compra (%)", x = "Grupos de
Tratamiento")+
theme(axis.text = element_text(size=12), axis.text.x = element_text(angle = 0),
text = element_text(size=12),
strip.text.x = element_text(size=12), legend.position = "bottom")+
facet_wrap(~descargo_app)
```





Pregunta 4:

¿Qué puedes concluir de la evaluación experimental? ¿Cuál sería tu recomendación para la empresa? ¿Vale la pena centrarse en un grupo específico de clientes?

Respuesta: Con base en la evaluación experimental, se concluye que ofrecer descuentos resulta más efectivo que proporcionar cashback para aumentar las compras y sus volúmenes. Aunque el impacto general del tratamiento es estadísticamente significativo, la presencia de subgrupos de clientes con características específicas indica que algunos son más sensibles que otros a realizar una compra cuando se les ofrece un descuento.

La recomendación para la empresa sería considerar una estrategia más focalizada, aprovechando la heterogeneidad en la respuesta de los clientes a las promociones. Identificar y dirigirse a esos subgrupos específicos que son más receptivos a los descuentos podría potenciar los beneficios y la rentabilidad de las promociones. Esto implica ajustar la estrategia de marketing y promociones para adaptarse a las preferencias y comportamientos de diferentes segmentos de clientes, maximizando así el impacto positivo en las conversiones y volúmenes de compra.

Avance 3

El objetivo de este avance es estimar un modelo de Causal Machine Learning para poder predecir el impacto de otorgar un descuento a nivel cliente. Es decir, la demanda incremental en las compras derivada de la promoción. Los resultados del modelo se utilizarán para implementar una estrategia de descuentos focalizada que maximice la rentabilidad.

Instalación de librerías

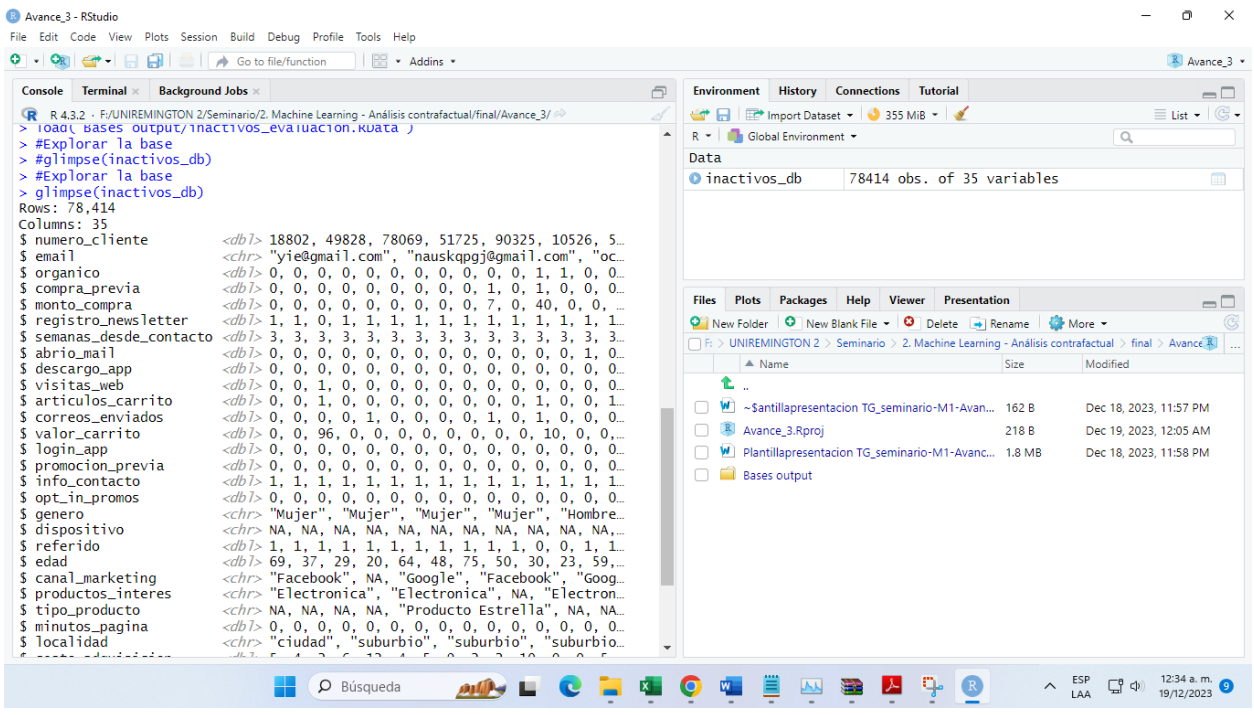
```
required_pkgs <- c('tidyverse', 'dplyr','RCT','grf',"fastDummies")
installed_pkgs <- installed.packages()
missing_pkgs <- required_pkgs[!(required_pkgs %in% installed_pkgs[, 1])]
if (length(missing_pkgs) == 0 ) {
message("Librerias cargadas")
} else {
```

```
install.packages(missing_pkgs)
message("Instalacion completa")
}
rm(installed_pkgs,missing_pkgs)
invisible(lapply(required_pkgs, library, character.only = TRUE))
rm(required_pkgs)
```

Pregunta 1

Explora la base y asegúrate de tener todas tus variables en formato numérico. Si tienes variables de texto o factores, transfórmalas a variables categóricas. Si tienes variables con valores vacíos decide si debes excluir a esas observaciones o mantenerlas y justifica tu decisión.

```
#Cargar la base
load("Bases output/inactivos_evaluacion.RData")
#Explorar la base
glimpse(inactivos_db)
```



```
#Generar la comparativa de medias
inactivos_db <- inactivos_db %>%
mutate(grupo_edad = ntile_label(edad,4,0))%>%
dummy_cols(select_columns = c("genero","dispositivo","canal_marketing",
"productos_interes","tipo_producto",
"localidad","grupo_edad"),
ignore_na = T, remove_selected_columns = T) %>%
mutate_at(vars(c(starts_with(c("genero","dispositivo","canal_marketing",
"productos_interes","tipo_producto","localidad")))),
function(x) x = if_else(is.na(x),0,as.double(x)))%>%
select(-c(email, edad, strata, missfit))

# Quitar caracteres especiales del nombre de las columnas
names(inactivos_db) <- make.names(names(inactivos_db))

# Analizar la distribucion de las variables
summary_stat <- summary_statistics(inactivos_db %>%
select(-c(numero_cliente, treat, treatment)))

# Winzorizar las variables con outliers
inactivos_db <- inactivos_db%>%
mutate_at(vars(monto_compra, valor_carrito,visitas_web,login_app),
function(x) x = if_else(x > quantile(x, probs = 0.99, na.rm = T),
quantile(x, probs = 0.99, na.rm = T), x))

# Valores faltantes
missings<-map_dbl(inactivos_db %>% select_all(),
~100*sum(is.na(.))/nrow(inactivos_db))
missings[missings>0]
## named numeric(0)
```

Pregunta 2:

Estima una matriz de correlaciones de todas tus variables. Muestra los pares de variables que tienen más de 95% de correlación y elimina una de cada par multicolineal.

```
# Construir la matriz de correlación
cor_matrix <- cor(inactivos_db %>%
select(-c(numero_cliente, treat, treatment)))
cor_matrix[upper.tri(cor_matrix, diag = T)] = NA
cor_tibble <- tibble(row = rep(rownames(cor_matrix), ncol(cor_matrix)),
col = rep(colnames(cor_matrix), each = ncol(cor_matrix)),
cor = as.vector(cor_matrix))
cor_tibble <- cor_tibble%>%filter(!is.na(cor))
large_cor_tibble <- cor_tibble%>%filter(abs(cor)>=0.95)

# Eliminar variables altamente correlacionadas
inactivos_db <- inactivos_db%>%select(-all_of(large_cor_tibble$col))

# Guardar la base de estimacion
save(inactivos_db, file = "Bases output/inactivos_estimacion.RData")
```

Pregunta 3.

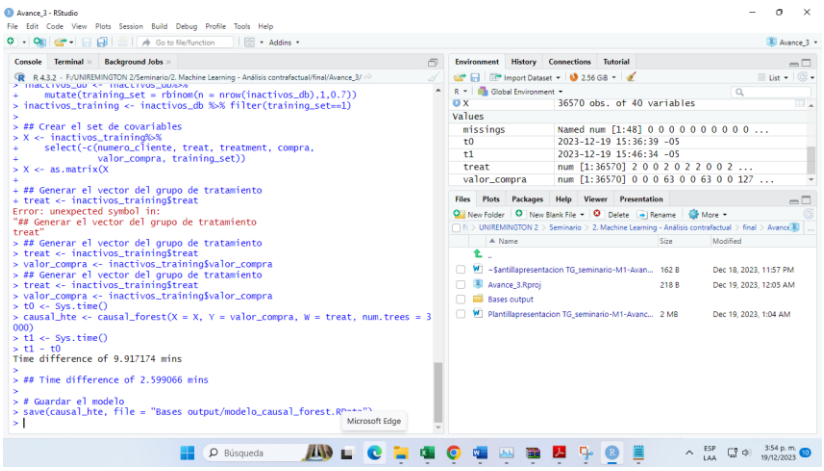
Quédate únicamente con tus clientes del grupo de control y el tratamiento con el descuento. Divide aleatoriamente a la población en 2 muestras: la muestra de entrenamiento (70% de las observaciones) y la muestra de validación (30% de las observaciones).

```
inactivos_db <- inactivos_db%>%
filter(treat != 1)
inactivos_db <- inactivos_db%>%
mutate(training_set = rbinom(n = nrow(inactivos_db),1,0.7))
inactivos_training <- inactivos_db %>% filter(training_set==1)
```

Pregunta 4:

Estima un causal forest en la base de entrenamiento (Estima 1000 árboles)

```
## Crear el set de covariables
X <- inactivos_training%>%
select(-c(numero_cliente, treat, treatment, compra,
valor_compra, training_set))
X <- as.matrix(X)

## Generar el vector del grupo de tratamiento
treat <- inactivos_training$treat
valor_compra <- inactivos_training$valor_compra
t0 <- Sys.time()
causal_hte <- causal_forest(X = X, Y = valor_compra, W = treat, num.trees = 3000)
t1 <- Sys.time()
t1 - t0

## Time difference of 2.599066 mins
# Guardar el modelo
save(causal_hte, file = "Bases output/modelo_causal_forest.RData")
```



Pregunta 5:

Realiza un histograma para mostrar la distribución del impacto de tratamiento ¿Cuál es el impacto promedio de tratamiento estimado por el modelo? ¿Se asimila al efecto encontrado en el experimento?

tau_in_sample = predict(causal_hte, estimate.variance = TRUE)

inactivos_training <- bind_cols(inactivos_training, tau_in_sample)

rm(tau_in_sample)

ggplot(inactivos_training)+

geom_histogram(aes(predictions),bins = 100,

fill = 'lightsteelblue', color = 'darkgrey')+

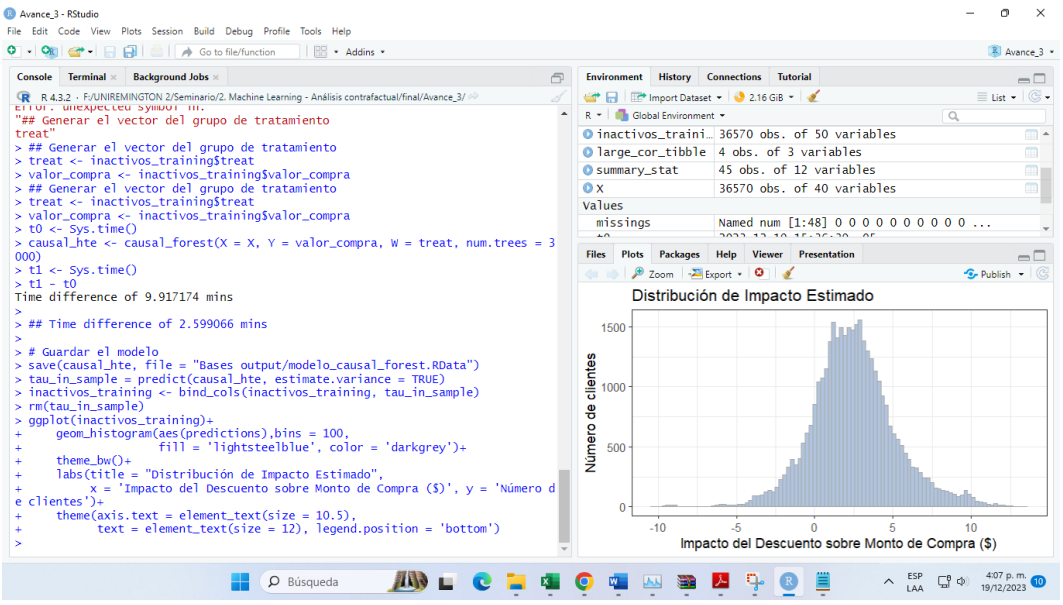
theme_bw()+

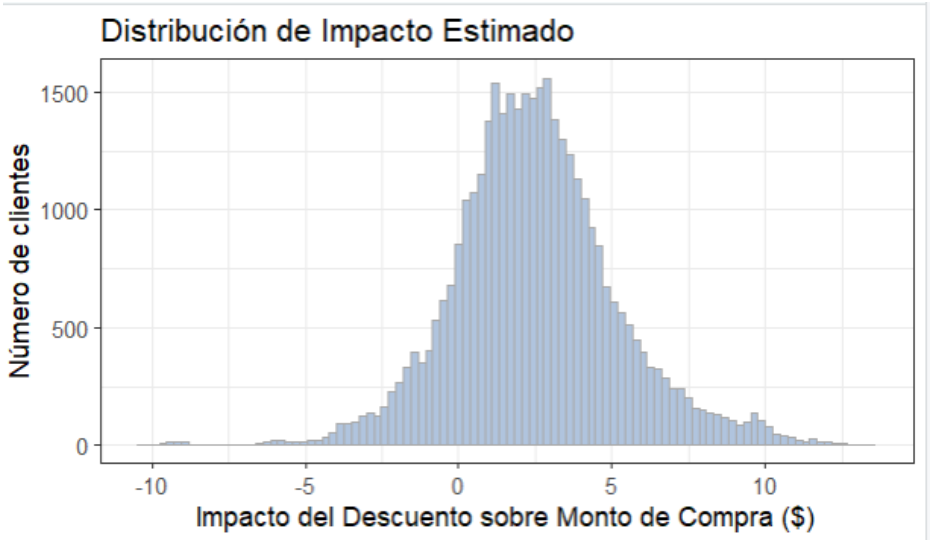
labs(title = "Distribución de Impacto Estimado",

x = 'Impacto del Descuento sobre Monto de Compra ($)', y = 'Número de clientes')+

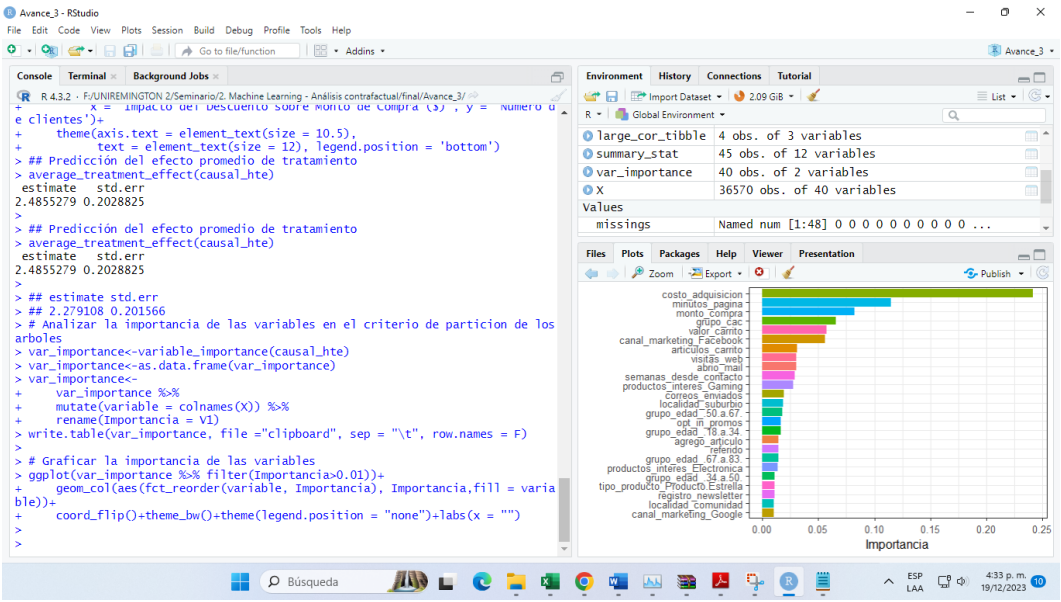
theme(axis.text = element_text(size = 10.5),

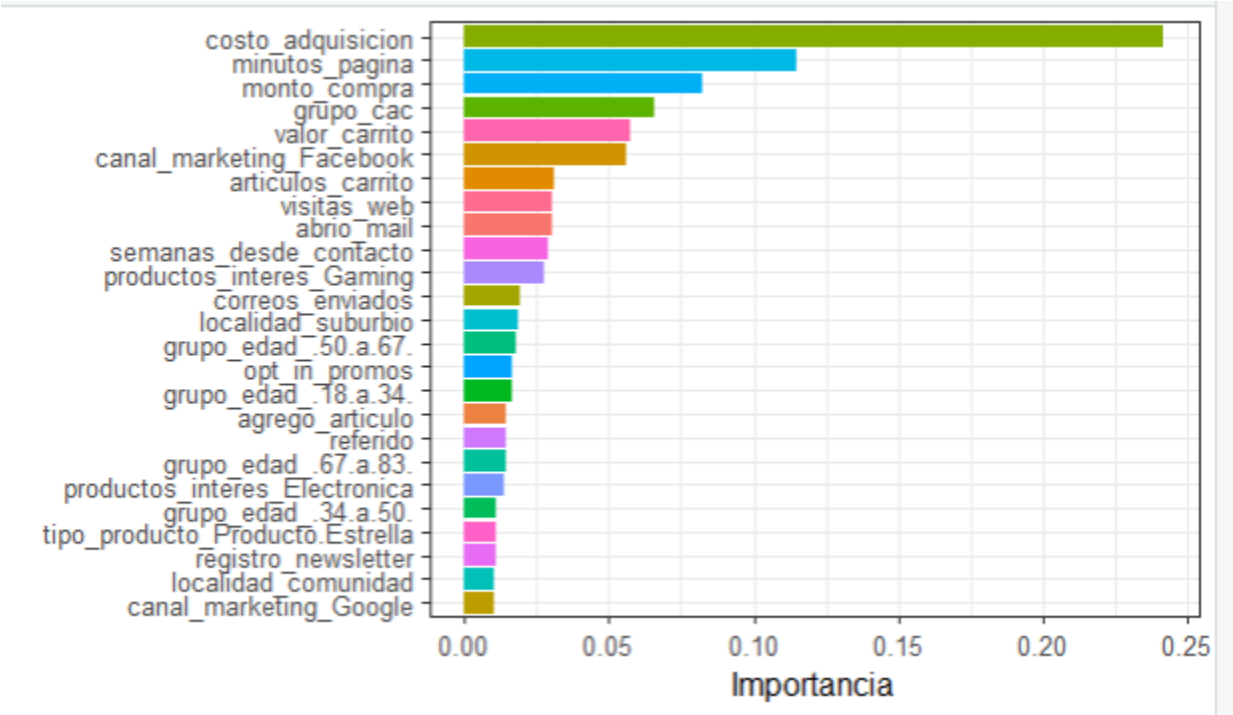
text = element_text(size = 12), legend.position = 'bottom')






```
## Predicción del efecto promedio de tratamiento
average_treatment_effect(causal_hte)

## estimate std.err
## 2.279108 0.201566

# Analizar la importancia de las variables en el criterio de particion de los arboles
var_importance<-variable_importance(causal_hte)
var_importance<-as.data.frame(var_importance)
var_importance<-
var_importance %>%
mutate(variable = colnames(X)) %>%
rename(Importancia = V1)
write.table(var_importance, file ="clipboard", sep = "\t", row.names = F)

# Graficar la importancia de las variables
ggplot(var_importance %>% filter(Importancia>0.01))+
geom_col(aes(fct_reorder(variable, Importancia), Importancia,fill = variable))+
coord_flip()+theme_bw()+theme(legend.position = "none")+labs(x = "")
```





rm(cor_matrix, cor_tibble, summary_stat, large_cor_tibble, X, var_importance, treat, valor_compra, missings)

Pregunta 6:

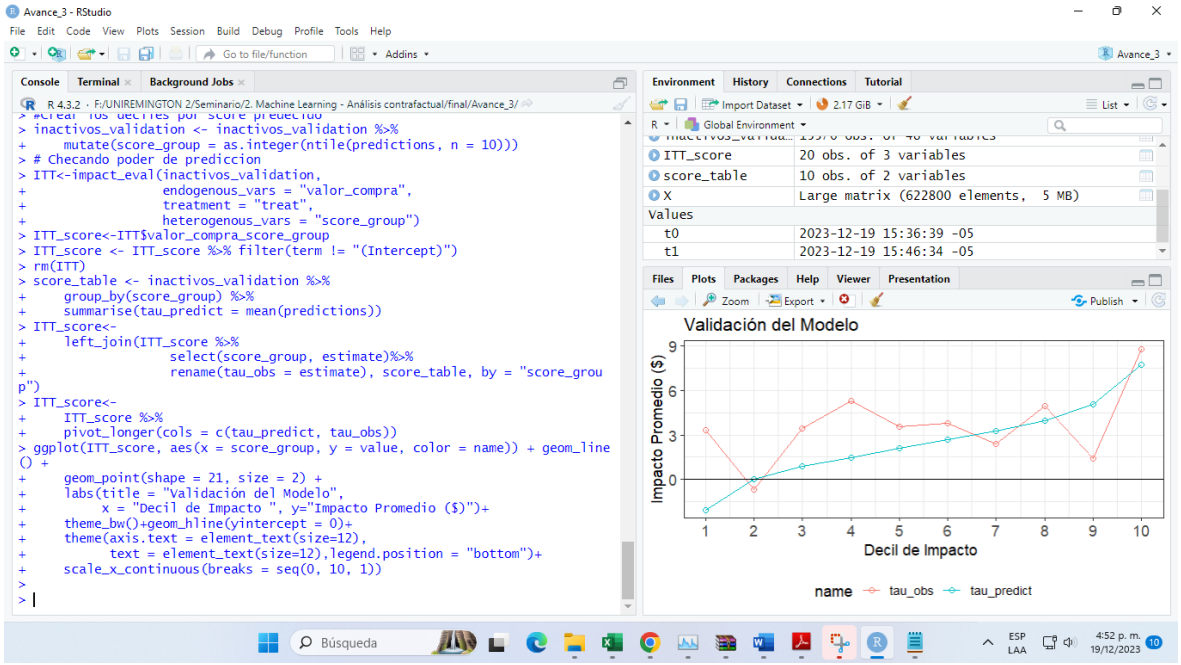
Evalua el poder predictivo del modelo en la base validacion. Recuerda dividir tu base de validación en k partes (k=10) con base en el score de predicción modelo. Posteriormente, estima el impacto de tratamiento (del experimento) en cada grupo de score y también calcula el promedio de las predicciones en cada grupo. (Tip: Puedes estimar el impacto de tratamiento con la función impact_eval considerando efectos heterogéneos por grupo de score). Valida si para los distintos grupos de score, la predicción del impacto promedio y el coeficiente de la regresión son crecientes y consistentes.

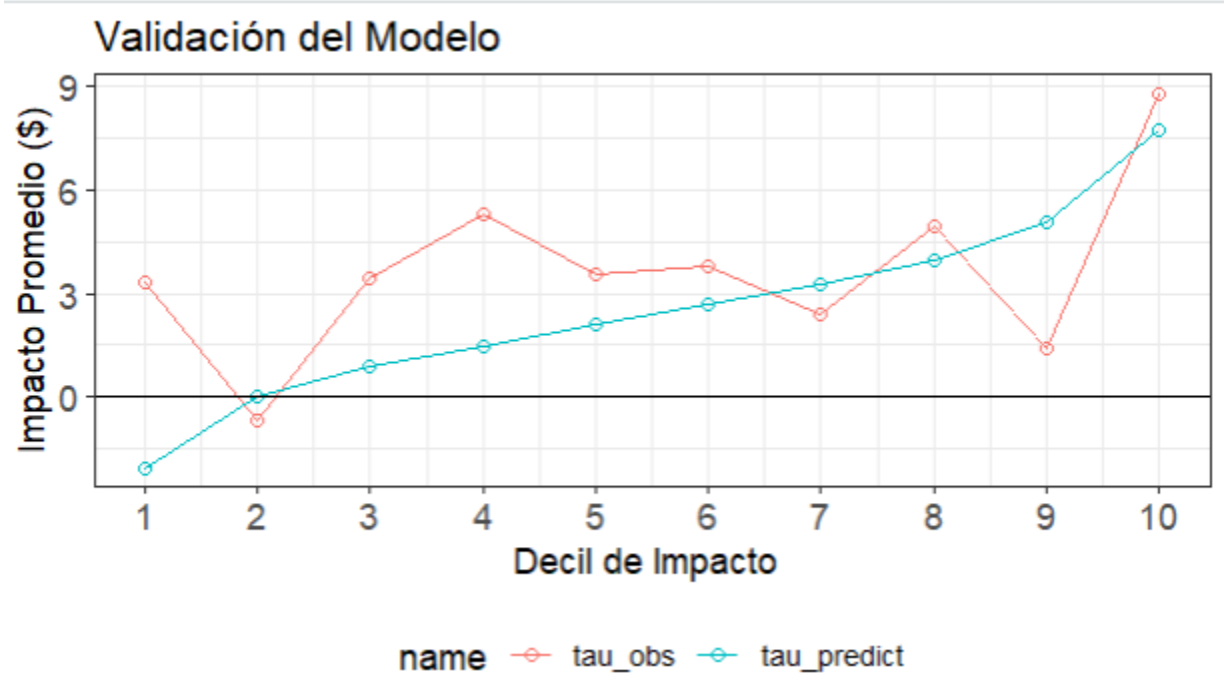
```
inactivos_validation <- inactivos_db %>% filter(training_set==0)

## Creamos el set de covariables de la validacion
X <- inactivos_validation%>%
select(-c(numero_cliente, treat, treatment, compra,
valor_compra, training_set))
X <- as.matrix(X)

# Realizamos la prediccion
inactivos_validation <- inactivos_validation %>%
mutate(predictions = predict(causal_hte, newdata = X)$predictions)
summary(inactivos_validation$predictions)
```

```
> # Realizamos la prediccion
> inactivos_validation <- inactivos_validation %>%
+     mutate(predictions = predict(causal_hte, newdata = X)$predictions)
> summary(inactivos_validation$predictions)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-9.1100  0.8884  2.3762  2.5005  3.9588 12.2641
```

```
## Min. 1st Qu. Median Mean 3rd Qu. Max.
## -10.1717 0.6605 2.1426 2.2687 3.7735 12.2707
```

```
#Crear los deciles por score predecido
inactivos_validation <- inactivos_validation %>%
mutate(score_group = as.integer(ntile(predictions, n = 10)))

# Checando poder de prediccion
ITT<-impact_eval(inactivos_validation,
endogenous_vars = "valor_compra",
treatment = "treat",
heterogenous_vars = "score_group")
ITT_score<-ITT$valor_compra_score_group
ITT_score <- ITT_score %>% filter(term != "(Intercept)")
rm(ITT)
score_table <- inactivos_validation %>%
group_by(score_group) %>%
summarise(tau_predict = mean(predictions))
ITT_score<-
left_join(ITT_score %>%
select(score_group, estimate)%>%
rename(tau_obs = estimate), score_table, by = "score_group")
ITT_score<-
ITT_score %>%
pivot_longer(cols = c(tau_predict, tau_obs))
ggplot(ITT_score, aes(x = score_group, y = value, color = name)) + geom_line() +
geom_point(shape = 21, size = 2) +
labs(title = "Validación del Modelo",
x = "Decil de Impacto ", y="Impacto Promedio ($)")+
theme_bw()+geom_hline(yintercept = 0)+
theme(axis.text = element_text(size=12),
text = element_text(size=12),legend.position = "bottom")+
scale_x_continuous(breaks = seq(0, 10, 1))
```


Avance_3 - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Console Terminal Background Jobs
R 4.3.2 · F:/UNIREMINGTON 2/Seminario/2. Machine Learning - Análisis contrafactual/final/Avance_3/
> inactivos_validation <- inactivos_validation %>%
+ mutate(score_group = as.integer(ntile(predictions, n = 10)))
> # Checando poder de prediccion
> ITT<-impact_eval(inactivos_validation,
+ endogenous_vars = "valor_compra",
+ treatment = "treat",
+ heterogenous_vars = "score_group")
> ITT_score<-ITT$valor_compra_score_group
> ITT_score <- ITT_score %>% filter(term != "(Intercept)")
> rm(ITT)
> score_table <- inactivos_validation %>%
+ group_by(score_group) %>%
+ summarise(tau_predict = mean(predictions))
> ITT_score<-
+ left_join(ITT_score %>%
+ select(score_group, estimate)%>%
+ rename(tau_obs = estimate), score_table, by = "score_grou
p")
> ITT_score<-
+ ITT_score %>%
+ pivot_longer(cols = c(tau_predict, tau_obs))
> ggplot(ITT_score, aes(x = score_group, y = value, color = name)) + geom_line
() +
+ geom_point(shape = 21, size = 2) +
+ labs(title = "Validación del Modelo",
+ x = "Decil de Impacto ", y="Impacto Promedio ($)")+
+ theme_bw()+geom_hline(yintercept = 0)+
+ theme(axis.text = element_text(size=12),
+ text = element_text(size=12),legend.position = "bottom")+
+ scale_x_continuous(breaks = seq(0, 10, 1))
Environment History Connections Tutorial
Import Dataset 2.17 GiB
Global Environment
ITT_score 20 obs. of 3 variables
score_table 10 obs. of 2 variables
X Large matrix (622800 elements, 5 MB)
Values
t0 2023-12-19 15:36:39 -05
t1 2023-12-19 15:46:34 -05
Files Plots Packages Help Viewer Presentation
Zoom Export Publish
Validación del Modelo
Impacto Promedio ($)
Decil de Impacto
name tau_obs tau_predict
Búsqueda
ESP LAA
4:52 p. m. 19/12/2023


Validación del Modelo
Impacto Promedio ($)
0
3
6
9
1
2
3
4
5
6
7
8
9
10
Decil de Impacto
name
tau_obs
tau_predict

Pregunta 7:

Predice cuál hubiera sido el impacto sobre las ventas si los clientes que recibieron el cashback hubieran recibido un descuento. Asume que estos clientes nunca fueron tratados y úsalos para simular una estrategia de focalización a nivel usuario con base en los resultados de tu modelo. Asume que el monto de compra mínimo es de $7 y que sólo tienes presupuesto para dar 1000 cupones de descuento ¿Cuál es el impacto promedio y el impacto total esperado de los usuarios de tu campaña focalizada?

```
load("Bases output/inactivos_estimacion.RData")
inactivos_focalizacion <- inactivos_db %>%
filter(treat==1)
# Generando el vector de covariables
X <- inactivos_focalizacion %>%
select(-c(numero_cliente, treat,
treatment, compra, valor_compra))

# Revisar que todas las covariables del modelo estén en la matrix
variables_modelo<-colnames(causal_hte$X.orig)
en_ambas<-intersect(variables_modelo, names(X))
variables_faltantes<-setdiff(variables_modelo, names(X))
X <- as.matrix(X)

# Realizar la prediccion
inactivos_focalizacion <- inactivos_focalizacion %>%
mutate(predictions = predict(causal_hte, newdata = X)$predictions)
summary(inactivos_focalizacion$predictions)
```

```
# Realizar la prediccion
inactivos_focalizacion <- inactivos_focalizacion %>%
    mutate(predictions = predict(causal_hte, newdata = X)$predictions)
summary(inactivos_focalizacion$predictions)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
9.1100  0.8898  2.3601  2.4876  3.9449 12.8994
```

Min. 1st Qu. Median Mean 3rd Qu. Max.

-10.1717 0.6651 2.1604 2.2639 3.7845 12.2707

Filtrar primeros 1000 clientes más responsivos

inactivos_focalizacion <- inactivos_focalizacion%>%

filter(rank(desc(predictions))<=1000 & predictions>7)

Impacto promedio e impacto total

lift_table <- inactivos_focalizacion%>%

group_by()%>%

summarise(impacto_promedio = mean(predictions),

impacto_esperado = impacto_promedio*1000*0.2080256)

Proyecto final de Introducción a la ética en la Inteligencia Artificial

Dilema ético

El proyecto final del curso, consiste en analizar y tratar de resolver un dilema ético relacionado a los conceptos de sesgos y prejuicios, los derechos de autor y la transparencia en la inteligencia artificial, similar al visto en la clase 4 del primer módulo del curso. El dilema es el siguiente:

Imagina una compañía tecnológica líder que desarrolla sistemas de IA para la industria de la salud. Estos sistemas son utilizados para ayudar a diagnosticar enfermedades y recomendar tratamientos. La compañía ha experimentado un éxito significativo en su tecnología y ha expandido su uso en hospitales y clínicas en todo el mundo.

Después de una revisión interna, la compañía descubre que su algoritmo tiene una tendencia a diagnosticar erróneamente ciertas condiciones médicas en personas de un grupo étnico específico. Este sesgo se debe a la falta de diversidad en los datos de entrenamiento. La corrección del algoritmo podría llevar tiempo y recursos significativos.

Además de esto, un grupo de pacientes descubre que sus datos y diagnósticos han sido utilizados para entrenar el algoritmo sin su consentimiento. Alegan que esto es una violación de sus derechos de privacidad.

Finalmente, hay crecientes preocupaciones entre los profesionales médicos y pacientes sobre cómo funciona exactamente la IA. Aunque la compañía ha declarado que su algoritmo es altamente preciso, no ha revelado completamente cómo llega a sus conclusiones. Esto ha llevado a una creciente desconfianza.
Con esto en mente, considera las siguientes preguntas para reflexionar sobre el caso ético:

Pasos

Para realizar este proyecto, debes seguir los siguientes pasos:

• Lee con detenimiento el dilema ético.

• Analiza cómo se pone en juego cada concepto ético en este caso hipotético.

El primer concepto ético observado en el caso es el de los sesgos y prejuicios en la IA. Esto constituye un problema, ya que el sesgo en el algoritmo afecta a un grupo étnico específico, lo que podría resultar en diagnósticos erróneos. Esto es un problema de tipo ético, porque la ausencia de diversidad en los datos de entrenamiento plantea la cuestión ética de la equidad y la no discriminación. Corregir este sesgo implica abordar la responsabilidad ética de garantizar que la IA sea justa y precisa para todos los grupos étnicos.

El otro concepto ético percibido en el caso es el de los derechos de autor y consentimiento, lo cual es un problema, porque hay un uso de datos y diagnósticos de pacientes sin su consentimiento, lo que representa una violación de la privacidad de los pacientes que plantea cuestiones éticas fundamentales. Por lo anterior, respetar los derechos de autor y obtener el consentimiento informado son principios éticos esenciales en el tratamiento de datos de salud.

Un tercer concepto ético tratado en el dilema es el de la transparencia en la IA. Este es un problema que nuestra claramente cómo el algoritmo tiene sus consecuencias. Es por eso, que la transparencia en el funcionamiento de la IA es crucial para la confianza de los usuarios y la rendición de cuentas, porque la falta de transparencia genera desconfianza y plantea preguntas éticas sobre la responsabilidad de la empresa.

Preguntas

• Resuelve el conflicto presentado respondiendo las siguientes preguntas (Recuerda basarte en lo que viste en este curso y tu perspectiva personal. No hay una única respuesta válida en cuanto a dilemas éticos.):

a. ¿Debería la compañía invertir en corregir el sesgo, incluso si eso significa un retraso en la implementación y posiblemente afectar su posición en el mercado?

Corregir el sesgo es fundamental desde una perspectiva ética para garantizar que la tecnología sea equitativa y no discriminadora, especialmente en el ámbito de la salud, donde la precisión es crucial para el bienestar de los pacientes. La compañía tiene la responsabilidad ética de ofrecer tecnologías seguras y justas. Ignorar o posponer la corrección del sesgo podría contribuir a disparidades en la atención médica y afectar negativamente a ciertos grupos étnicos. Además, la precisión en el diagnóstico es esencial en la atención médica. No abordar el sesgo podría tener consecuencias graves para la salud pública y la confianza en la tecnología.

Por otro lado, la transparencia y la responsabilidad en la corrección del sesgo pueden mejorar la reputación de la compañía a largo plazo. El compromiso con la equidad y la corrección activa de errores éticos puede generar confianza entre usuarios y partes interesadas. Asimismo, corregir el sesgo puede llevar tiempo y recursos, lo que podría resultar en un retraso en la implementación. Esto podría afectar la capacidad de la empresa para mantener o mejorar su posición en el mercado. También afecta, la competitividad, ya que, en un mercado competitivo, la velocidad de implementación es a menudo crucial. Un retraso significativo podría permitir que competidores aprovechen la oportunidad, especialmente si ya cuentan con sistemas que abordan estos problemas éticos.

En resumen, considerando la importancia ética y social de corregir el sesgo en un algoritmo de diagnóstico médico, se podría recomendar que la compañía invierta en la corrección del sesgo, incluso si esto implica un retraso en la implementación. Sin embargo, la estrategia de comunicación y gestión del tiempo durante este proceso es clave para minimizar el impacto en la posición en el mercado. Además, la compañía debería demostrar transparencia y compromiso con la equidad para mitigar posibles efectos negativos en su reputación. La prioridad debería ser la entrega de tecnologías que sean éticas, precisas y justas en el ámbito de la salud.

b. Si la compañía decide corregir el sesgo, es posible que mucha gente no logre obtener un diagnóstico de manera veloz o precisa, ¿vale la pena corregir el sesgo incluso si afecta negativamente a las personas que no se veían afectadas anteriormente?

La decisión de corregir el sesgo en el algoritmo de la compañía, a pesar de la posibilidad de afectar la velocidad y precisión en el diagnóstico para algunas personas, involucra consideraciones éticas complejas. La corrección del sesgo es esencial para garantizar la equidad y la justicia en el acceso a la atención médica. Ignorar el sesgo podría perpetuar disparidades y discriminación en la atención médica, lo cual sería éticamente inaceptable.

La empresa tiene una responsabilidad social de proporcionar tecnologías seguras y justas que no pongan en peligro la salud o bienestar de ciertos grupos de personas. Por otro lado, la corrección del sesgo podría llevar a un aumento en el tiempo necesario para realizar diagnósticos precisos. Esto podría afectar a las personas que necesitan resultados rápidos, como en situaciones de emergencia. En cambio, no corregir el sesgo podría tener consecuencias graves para la salud pública, ya que podría llevar a diagnósticos incorrectos y tratamientos inadecuados, afectando negativamente a un grupo específico de personas.

En mi opinión, sí vale la pena corregir el sesgo incluso si afecta negativamente a las personas que no se veían afectadas anteriormente. La equidad y la justicia en la atención médica son principios éticos fundamentales, y la corrección del sesgo es crucial para cumplir con estos principios. Sin embargo, la compañía debería adoptar un enfoque proactivo para minimizar cualquier impacto negativo en la velocidad del diagnóstico. Esto podría incluir estrategias como el desarrollo acelerado de soluciones temporales mientras se trabaja en una corrección a largo plazo, o la implementación de procesos eficientes para situaciones críticas. Además, la empresa debe comunicar claramente a los profesionales médicos y a los usuarios sobre los cambios, brindando información transparente sobre el proceso de corrección y los posibles impactos en la velocidad del diagnóstico. Esto contribuiría a la confianza del público y demostraría un compromiso continuo con la equidad y la justicia en el desarrollo de tecnologías de inteligencia artificial.

c. ¿Cómo debería la compañía abordar las acusaciones de infracción de violación de la privacidad? ¿Deberían compensar a los pacientes cuyos datos se utilizaron sin permiso, y cómo podría hacerse esto sin crear un precedente peligroso?
¿Cómo debería la compañía abordar las acusaciones de infracción de violación de la privacidad?
La compañía debería abordar las acusaciones de infracción de violación de la privacidad de la siguiente forma:

La compañía debe reconocer públicamente la infracción de privacidad y asumir la responsabilidad por el uso no autorizado de los datos de los pacientes.
También debe llevar a cabo una investigación interna para comprender completamente el alcance de la infracción, identificar cómo ocurrió y quiénes fueron afectados. Además, la compañía debe considerar la compensación a los pacientes cuyos datos se utilizaron sin permiso. Esto podría incluir opciones como asistencia médica gratuita, servicios adicionales de salud o incluso compensaciones financieras. Otra gran idea sería que implemente medidas adicionales para fortalecer la seguridad y la privacidad de los datos de los pacientes, asegurando que no se repita la violación en el futuro. de igual modo, debería revisar y mejorar las prácticas de obtención de consentimiento para garantizar que los pacientes estén completamente informados sobre cómo se utilizarán sus datos y tengan la opción de dar o retirar su consentimiento. Otro aspecto que la compañía no debe pasar por alto es comunicar de manera transparente a los pacientes y al público en general sobre las acciones tomadas para abordar la violación de privacidad y las medidas implementadas para prevenir futuros incidentes. También sería oportuno que colabore con las autoridades regulatorias pertinentes para garantizar que la compañía cumpla con las leyes y regulaciones de privacidad.

¿Deberían compensar a los pacientes cuyos datos se utilizaron sin permiso, y cómo podría hacerse esto sin crear un precedente peligroso?

La compensación a los pacientes debería abordarse como una solución específica para este caso en particular, evitando establecer un precedente general que pueda ser explotado en situaciones no comparables. Del mismo modo, debería implementar medidas sólidas para proteger los datos en el futuro y evitar violaciones recurrentes, demostrando así un compromiso continuo con la privacidad. Además, es importante que sean transparencia en las condiciones de compensación: Ser transparentes sobre las condiciones de compensación, de manera que no se perciba como un intento de evitar responsabilidades, sino como un esfuerzo genuino para remediar los daños causados. Por

último, la compañía debería mantener un diálogo continuo con los pacientes afectados y considerar sus necesidades y preocupaciones al abordar la compensación.
En resumen, la compañía debe abordar las acusaciones con transparencia, asumir responsabilidad, mejorar las prácticas de privacidad y considerar medidas de compensación específicas para los pacientes afectados, todo ello con el objetivo de restablecer la confianza y preservar la integridad ética de la empresa.

d. ¿Hasta qué punto debería la compañía revelar cómo funciona su algoritmo?

La transparencia en los algoritmos de inteligencia artificial (IA) es crucial para construir confianza. La compañía debe equilibrar la transparencia con la protección de propiedad intelectual y seguridad. Consideraciones clave incluyen proporcionar explicaciones claras, proteger la propiedad intelectual, permitir auditorías independientes y adoptar una transparencia selectiva. Facilitar la educación pública, cumplir con regulaciones éticas y mantener un diálogo continuo con partes interesadas también son aspectos esenciales. En última instancia, encontrar un equilibrio entre transparencia y seguridad es crucial para mantener la confianza del público y adaptarse a las cambiantes expectativas y regulaciones.

e. ¿Cómo pueden los líderes de la compañía equilibrar las demandas de todas estas áreas y asegurar que su tecnología no solo sea comercialmente exitosa sino también éticamente responsable?

Para lograr un equilibrio entre las demandas comerciales y éticas, los líderes deben implementar estrategias clave, como fomentar una cultura ética, integrar consideraciones éticas desde las etapas iniciales del desarrollo de productos y proporcionar formación en ética a los equipos. Además, establecer políticas de transparencia, realizar auditorías éticas y permitir revisiones externas contribuyen a una evaluación independiente. La participación activa de partes interesadas, el cumplimiento de normativas y estándares éticos, y la creación de mecanismos de rendición de cuentas son esenciales. Asimismo, la evaluación de impacto ético y una comunicación transparente con usuarios y partes interesadas refuerzan el compromiso con la responsabilidad social. El liderazgo con

integridad y la adaptación constante a las cambiantes expectativas y tecnologías son clave para mantener el equilibrio entre el éxito comercial y la responsabilidad ética.

Proyecto final de Innovación tecnológica con inteligencia artificial

1. Base de datos elegida.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import scipy.special

pd.set_option('display.float_format', lambda x: '%.2f' % x)

test = pd.read_csv("ventas-por-factura.csv")
df = test.copy()
df.head()
```

```
Requirement already satisfied: Lifetimes==0.2.2.2 in c:\users\mauro\anaconda3\lib\site-packages (0.2.2.2)
Requirement already satisfied: numpy in c:\users\mauro\anaconda3\lib\site-packages (from Lifetimes==0.2.2.2) (1.24.3)
Requirement already satisfied: scipy in c:\users\mauro\anaconda3\lib\site-packages (from Lifetimes==0.2.2.2) (1.11.1)
Requirement already satisfied: pandas>=0.19 in c:\users\mauro\anaconda3\lib\site-packages (from Lifetimes==0.2.2.2) (2.0.3)
Requirement already satisfied: python-dateutil>=2.8.2 in c:\users\mauro\anaconda3\lib\site-packages (from pandas>=0.19->Lifetim
es==0.2.2.2) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in c:\users\mauro\anaconda3\lib\site-packages (from pandas>=0.19->Lifetimes==0.2.2.
2) (2023.3.post1)
Requirement already satisfied: tzdata>=2022.1 in c:\users\mauro\anaconda3\lib\site-packages (from pandas>=0.19->Lifetimes==0.2.
2.2) (2023.3)
Requirement already satisfied: six>=1.5 in c:\users\mauro\anaconda3\lib\site-packages (from python-dateutil>=2.8.2->pandas>=0.1
9->Lifetimes==0.2.2.2) (1.16.0)
```

Out[2]:

	N° de factura	Fecha de factura	ID Cliente	País	Cantidad	Monto
0	548370	3/30/2021 16:14:00	15528.00	United Kingdom	123	229,33
1	575767	11/11/2021 11:11:00	17348.00	United Kingdom	163	209,73
2	C570727	10/12/2021 11:32:00	12471.00	Germany	-1	-1,45
3	549106	4/6/2021 12:08:00	17045.00	United Kingdom	1	39,95
4	573112	10/27/2021 15:33:00	16416.00	United Kingdom	357	344,83

2. Preguntas de investigación.

¿Cuál es el patrón de ventas mensual y las tendencias estacionales basándose en la cantidad y el monto de las transacciones registradas en la base de datos?

¿Cuál es el país que contribuye con el mayor volumen de ventas y cuál tiene el mayor valor monetario en términos de transacciones registradas?

¿Hay algún cliente específico que destaque por su frecuencia de compra o por el monto total gastado en las transacciones?

¿Existe alguna relación significativa entre la cantidad de productos comprados y el monto total de la factura? ¿Esta relación varía según el país o el cliente?

¿Cómo afectan las devoluciones (transacciones con cantidad negativa) al desempeño financiero general de la empresa y qué productos o clientes están más asociados con devoluciones?

3. Pregunta seleccionada y columnas parametrizadas.

Pregunta Seleccionada: ¿Hay algún cliente específico que destaque por su frecuencia de compra o por el monto total gastado en las transacciones?

Columnas parametrizadas:

ID Cliente: Para identificar y agrupar las transacciones por cliente.

Cantidad: Para calcular la frecuencia de compra por cliente.

Monto: Para evaluar el monto total gastado por cada cliente en todas sus transacciones.

4. Pasos a seguir y elección de IA.

Pasos:

Exploración de Datos:

Examina la distribución de la frecuencia de compra y el monto total gastado por cliente. Identifica valores atípicos y patrones notables.

Agrupación por Cliente:

Agrupar las transacciones por ID de cliente para calcular la frecuencia de compra y el monto total gastado por cada uno.

Análisis Estadístico Descriptivo:

Calcular estadísticas descriptivas, como la media, mediana, y desviación estándar, para comprender la tendencia central y la dispersión de los datos.

Visualización de Datos:

Utilizar gráficos y visualizaciones para representar la frecuencia de compra y el monto total gastado por cliente.

Selección de Modelo de Inteligencia Artificial:

Dado que se está trabajando con datos de clientes y se quiere identificar patrones, un modelo de clustering podría ser apropiado. El algoritmo K-Means es una opción común para agrupar clientes en función de su comportamiento de compra.

Entrenamiento del Modelo:

Preparar los datos, se debe elegir el número óptimo de clústeres, y entrenar el modelo K-Means con la información de frecuencia de compra y monto total gastado.

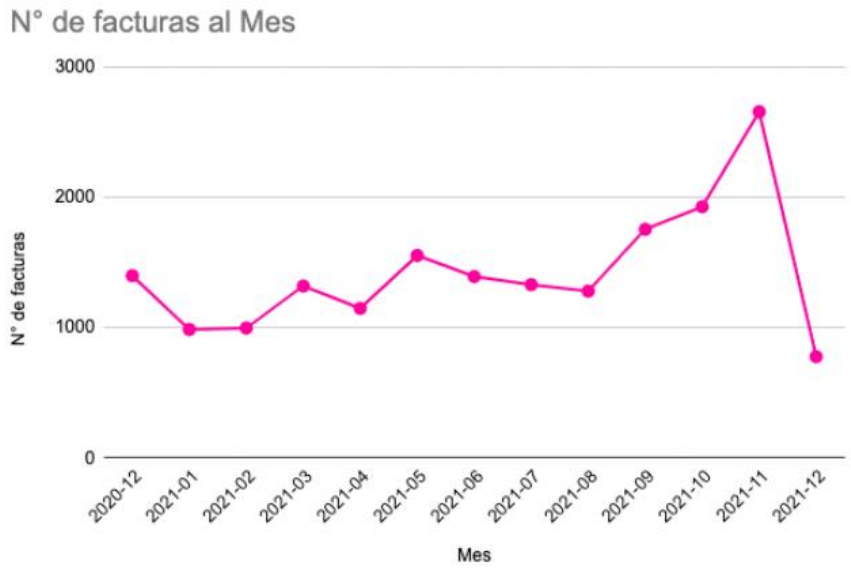
Evaluación del Modelo:

Se debe evaluar la calidad de los clústeres generados por el modelo para determinar si hay clientes específicos que destacan en términos de frecuencia de compra o monto total gastado.

Interpretación de Resultados:

Analizar los clústeres identificados y determinar si hay grupos de clientes que se destacan en términos de comportamiento de compra.

5. Representación gráfica de los resultados.



Proyecto final de Introducción a la Inteligencia Artificial

Parte 1: Crea tu modelo de 3 categorías

Para esta primera entrega del proyecto deberás aplicar lo visto en el módulo de visión por computadora, los pasos que deberás seguir son exactamente los mismos que vimos en clase, sólo tendrás que agregar una clase adicional a la clasificación, entrenar el modelo y ejecutarlo desde Google Colab.

Para lograr este primer avance, debes seguir estos pasos:

1. Abre la página de http://teachablemachine.withgoogle.com
2. Selecciona objetos para generar 3 clases que el modelo va a ser capaz de identificar
3. Entrena tu modelo
4. Exporta el modelo y copia y pega el código generado con tensorflow en google colab.
5. Prueba el modelo con una nueva foto de alguna de las 3 clases de las que entrenaste el modelo.
6. Genera la predicción.
7. Sube el archivo de Google Colab a la plataforma de Crehana.

```
from google.colab import drive
drive.mount('/content/drive')

import tensorflow.keras
from PIL import Image, ImageOps
import numpy as np

from tensorflow.keras.optimizers import RMSprop
from tensorflow.keras.preprocessing.image import ImageDataGenerator

np.set_printoptions(suppress=True)
```

Sustituye el modelo con el que exportaste:

```
model = tensorflow.keras.models.load_model('/content/keras_model.h5')

data = np.ndarray(shape=(1, 224, 224, 3), dtype=np.float32)
```

Carga la imagen a analizar:

```
image = Image.open('/content/pelota.jpg')
image = image.convert('RGB')

size = (224,224)
image = ImageOps.fit(image,size, Image.ANTIALIAS)

image_array = np.asarray(image)

image.show()

normalized_image_array = (image_array.astype(np.float32)/127.0) -1

data[0].shape

data[0].ndim

data[0] = normalized_image_array

prediction = model.predict(data)

prediction

np.argmax(prediction)
```

Parte 2: Tu proyecto final

El segundo avance de tu proyecto final corresponde al desarrollo de una red neuronal, en la cual utilizaremos un nuevo set de datos llamado CIFAR100. En este avance vas a tener que redimensionar los datos, construir tu red neuronal para entrenarlo y luego hacer la evaluación.

Lo que tendrás que hacer es: 1. Abrir la plantilla “Plantilla_NN_ CIFAR100.ipynb” desde Google Colab donde encontrarás el paso a paso para desarrollar el proyecto.

2. Lo primero será importar las librerías necesarias para importar los datos y transformarlos.

3. Después, deberás revisar cuidadosamente las dimensiones de tu dataset para poder hacer las transformaciones necesarias para que las reciba la red.

4. Deberás crear la red neuronal con esas nuevas dimensiones.

5. Genera un set de datos para train y test con el set de datos originales.

6. Convierte las variables a variables categóricas para poderlas ingresar al modelo.

7. Entrena la red neuronal.

8. Genera una evaluación para train y test.

9. Finalmente, sube el archivo de Google Colab a la plataforma de Crehana.

```
# Importando librerías

import numpy as np
import matplotlib.pyplot as plt

# Carga de Datos
```

Importamos las librerías que utilziaremos para cargar los datos para la red neuronal:

```
from keras.datasets import mnist
from keras import layers, models


(train_data, train_labels), (test_data, test_labels) = mnist.load_data()

train_data.shape

train_data[0]

plt.imshow(train_data[0])
plt.show

train_labels[0]

# Modelo

model = models.Sequential()
model.add(layers.Dense(512, activation='relu', input_shape=(28*28,))) #Tenemos 512
neuronas de entrada, que tienen una forma de 28*28 px
model.add(layers.Dense(10, activation = 'softmax')) #10 neuronas de la posible salida
```

```
model.compile(optimizer='rmsprop',
              loss = 'categorical_crossentropy', #Función de pérdida
              metrics = ['accuracy'] #Variable a optimizar
)

# Transformación de datos

x_train = train_data.reshape((60000, 28*28))
x_train = x_train.astype('float32')/255

x_test = test_data.reshape(10000, 28*28)
x_test = x_test.astype('float32')/255

from tensorflow.keras.utils import to_categorical

y_train = to_categorical(train_labels)
y_test = to_categorical(test_labels)

train_labels[0]

y_train[0]

# Entrenamiento

Iteraciones de las épocas

model.fit(x_train, y_train, epochs=5, batch_size=128)

# Evaluación
```

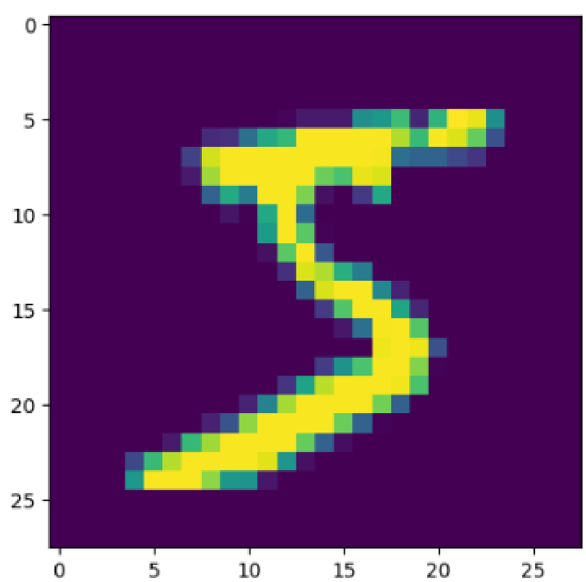
model.evaluate(x_test, y_test)

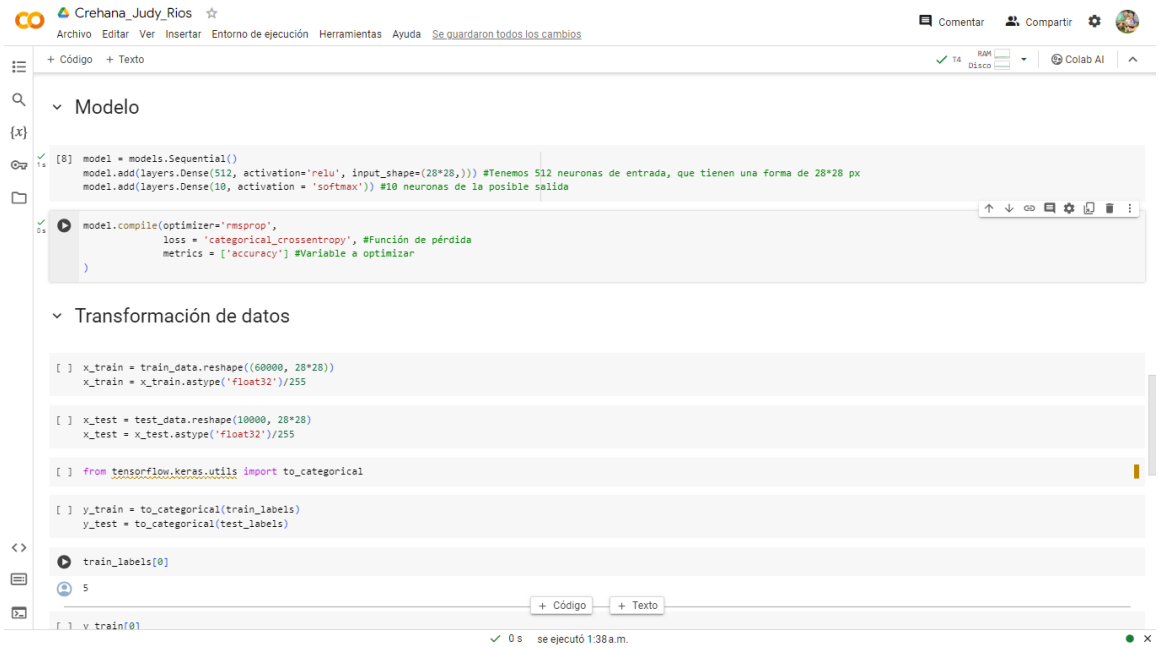
```
[3]
    Downloading data from https://storage.googleapis.com/tensorflow/tf-keras-datasets/mnist.npz
    11490434/11490434 [==============================] - 0s 0us/step
```

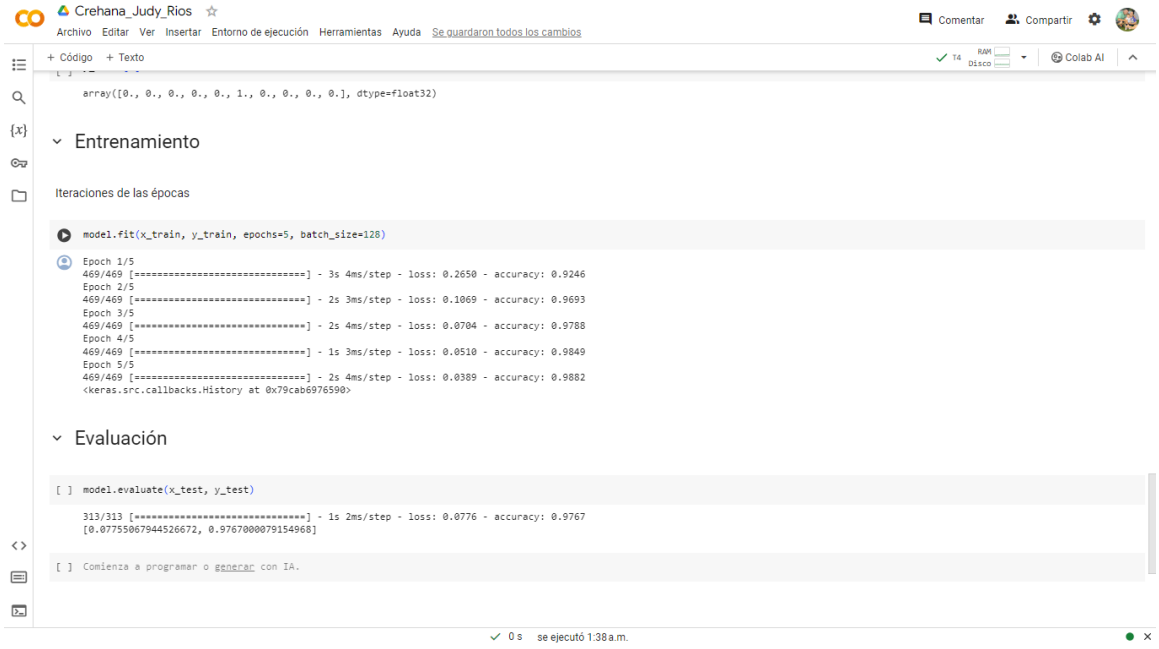
```
[4] train_data.shape

    (60000, 28, 28)
```

```
train_data[0]

array([[ 0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,
         0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,
         0,   0],
       [ 0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,
         0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,
         0,   0],
       [ 0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,
         0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,
         0,   0],
       [ 0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,
         0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,
         0,   0],
       [ 0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,
         0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
```

```
plt.imshow(train_data[0])
plt.show

<function matplotlib.pyplot.show(close=None, block=None)>
```




Crehana_Judy_Rios
Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Se guardaron todos los cambios
Comentar Compartir
+ Código + Texto
Colab AI
Modelo
[8] model = models.Sequential()
model.add(layers.Dense(512, activation='relu', input_shape=(28*28,))) #Tenemos 512 neuronas de entrada, que tienen una forma de 28*28 px
model.add(layers.Dense(10, activation = 'softmax')) #10 neuronas de la posible salida
model.compile(optimizer='rmsprop',
loss = 'categorical_crossentropy', #Función de pérdida
metrics = ['accuracy'] #Variable a optimizar
)
Transformación de datos
x_train = train_data.reshape((60000, 28*28))
x_train = x_train.astype('float32')/255
x_test = test_data.reshape(10000, 28*28)
x_test = x_test.astype('float32')/255
from tensorflow.keras.utils import to_categorical
y_train = to_categorical(train_labels)
y_test = to_categorical(test_labels)
train_labels[0]
5
+ Código + Texto
0 s se ejecutó 1:38 a.m.


Crehana_Judy_Rios
Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Se guardaron todos los cambios
Comentar Compartir
+ Código + Texto
Colab AI
array([0., 0., 0., 0., 0., 1., 0., 0., 0., 0.], dtype=float32)
Entrenamiento
Iteraciones de las épocas
model.fit(x_train, y_train, epochs=5, batch_size=128)
Epoch 1/5
469/469 [==========] - 3s 4ms/step - loss: 0.2650 - accuracy: 0.9246
Epoch 2/5
469/469 [==========] - 2s 3ms/step - loss: 0.1069 - accuracy: 0.9693
Epoch 3/5
469/469 [==========] - 2s 4ms/step - loss: 0.0704 - accuracy: 0.9788
Epoch 4/5
469/469 [==========] - 1s 3ms/step - loss: 0.0510 - accuracy: 0.9849
Epoch 5/5
469/469 [==========] - 2s 4ms/step - loss: 0.0389 - accuracy: 0.9882
<keras.src.callbacks.History at 0x79cab6976590>
Evaluación
model.evaluate(x_test, y_test)
313/313 [==========] - 1s 2ms/step - loss: 0.0776 - accuracy: 0.9767
[0.07755067944526672, 0.9767000079154968]
Comienza a programar o generar con IA.
0 s se ejecutó 1:38 a.m.

Parte 3: Análisis de discurso

Para este último avance, utilizarás el archivo de datos de Spotify tendrás que realizar un análisis de discurso haciendo procesamiento del lenguaje natural, como lo visto en clase. De esta forma generarás una nube de palabras con los tópicos más frecuentes.

Para extraer los tweets sobre Spotify y obtener la información más relevante debes hacer lo siguiente:

1. Abre la plantilla “Plantilla_PLN_Spotify.ipynb” desde Google Colab.
2. Deberás importar el archivo csv con la información sobre Spotify que se encuentra en los archivos adjuntos.
3. Sigue cada uno de los pasos que vienen en el notebook: Hacer las transformaciones a la información, limpiar los datos, generar los tokens y luego convertirlos a vectores para trabajar con ellos.
4. Una vez realizado esto tendrás que ejecutar el flujo de procesamiento de lenguaje natural para extraer la frecuencia de las palabras.
5. Finalmente con esto podrás generar una nube de palabras y con ellas obtener los insights más relevantes de los tópicos alrededor de las palabras claves.
6. Con esto terminado ya podrás subir el archivo de Google Colab a la plataforma de Crehana.

[3]
```
consumer_key = ''
consumer_secret = ''
access_token = ''
access_token_secret = ''
```

[67]
```
import tweepy
import pandas as pd
import datetime
```

[68]
```
auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)
```

[69]
```
api = tweepy.API(auth)
```