Entrega_Final_PP

Pablo Proaño

2025-12-01

PROYECTO FINAL

Analítica de Datos Industriales para la Toma de Decisiones con Apoyo de AI (Curso ADI&TD-IA)

Datos del Estudiante:

Nombre: Pablo Proaño

mail: pablo.proano@epn.edu.ec

Telf: +593 98 244 0935

Descripción del Proyecto

El presente proyecto tiene como objetivo analizar un conjunto de datos correspondiente a operaciones de importación registradas por el Servicio Nacional de Aduana del Ecuador (SENAE), específicamente del conjunto denominado SENAE_Importaciones Régimen General, el cual contiene el detalle de las importaciones por fecha de ingreso de la Declaración Aduanera de Importación (DAI) con estado de salida autorizada. La base de datos, disponible en formato CSV y actualizada a diciembre de 2025, es utilizada para realizar procesos de carga, limpieza, descripción estadística y visualización de variables relacionadas con valor, peso, cantidades e impuestos.

Descripción de la Base de Datos

La base de datos utilizada en este proyecto corresponde al conjunto SENAE_Importaciones Régimen General, publicado por el Servicio Nacional de Aduana del Ecuador (SENAE), y contiene información detallada de las operaciones de importación registradas mediante la Declaración Aduanera de Importación (DAI) con estado de salida autorizada. El conjunto de datos se presenta en formato CSV y está compuesto por variables de tipo categórico y numérico.

Objetivo del Proyecto

Para el desarrollo del proyecto se consideran como **variables de entrada** aquellas asociadas a la información conocida al arribo de la mercancía a Aduana, tales como SUBPARTIDA, PAIS_ORIGEN, REGIMEN, PESO_NETO, FOB, FLETE, SEGURO, CANTIDAD_FISICA y CANTIDAD_COMERCIAL. **Como variables de salida** se definen los principales tributos calculados por la autoridad aduanera, entre ellos ADVALOREM, ICE, FODINFA, SALVAGUARDIA e IVA. Si bien estos valores se determinan mediante

ecuaciones, formulas , tablas arancelarias y normativa específica, el objetivo del proyecto es **desarrollar un modelo basado en datos** que permita estimar dichos tributos aun sin conocer explícitamente estas ecuaciones, utilizando únicamente la información histórica contenida en la base de datos.

Carga y tratamiento de Datos

En esta sección cargamos los datos de la base que conseguimos para este trabajo.

```
library(readr)
aduana  <- read_delim("https://raw.githubusercontent.com/PjosP/Entrega_Final_Pablo_Proano/refs/heads/ma
    delim = "|", escape_double = FALSE, trim_ws = TRUE)
```

```
## Rows: 133493 Columns: 30
## -- Column specification --------------------------------------------------------
## Delimiter: "|"
## chr  (11): TIPO_IMPORTACION, ESTADO_DECLARACION, DISTRITO, REGIMEN, DESCRIPC...
## dbl  (18): SUBPARTIDA, CODIGO_COMPLEMENTARIO, CODIGO_SUPLEMENTARIO, PESO_NET...
## date  (1): FEC_INGRESO
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(aduana)
```

```
## # A tibble: 6 x 30
##   TIPO_IMPORTACION ESTADO_DECLARACION   FEC_INGRESO DISTRITO  REGIMEN SUBPARTIDA
##   <chr>            <chr>                <date>      <chr>     <chr>        <dbl>
## 1 IMP.GRAL.        10-SALIDA AUTORIZADA 2025-04-01  028-GUAY~ 10-IMP~ 9032100000
## 2 IMP.GRAL.        10-SALIDA AUTORIZADA 2025-09-08  028-GUAY~ 10-IMP~ 8544300000
## 3 IMP.GRAL.        10-SALIDA AUTORIZADA 2025-04-25  028-GUAY~ 10-IMP~ 8215200000
## 4 IMP.GRAL.        10-SALIDA AUTORIZADA 2025-06-03  055-QUITO 10-IMP~ 9617000000
## 5 IMP.GRAL.        10-SALIDA AUTORIZADA 2025-06-22  055-QUITO 10-IMP~ 6202900000
## 6 IMP.GRAL.        10-SALIDA AUTORIZADA 2025-02-21  028-GUAY~ 70-DEP~ 9505900000
## # i 24 more variables: CODIGO_COMPLEMENTARIO <dbl>, CODIGO_SUPLEMENTARIO <dbl>,
## #   DESCRIPCION_ARANCELARIA <chr>, PAIS_ORIGEN <chr>, CODIGO_LIBERACION <chr>,
## #   TRATAMIENTO_PREFERENCIAL <chr>, CONVENIO_INTERNACIONAL <chr>,
## #   TIPO_UNIDAD_FISICA <chr>, TIPO_UNIDAD_COMERCIAL <chr>, PESO_NETO <dbl>,
## #   FOB <dbl>, FLETE <dbl>, SEGURO <dbl>, CIF <dbl>, CANTIDAD_FISICA <dbl>,
## #   CANTIDAD_COMERCIAL <dbl>, SALVAGUARDIA <dbl>,
## #   SALVAGUARDIA_ESPECIFICA <dbl>, FODINFA <dbl>, ICE_ADVALOREM <dbl>, ...
```

Vamos a revisar las dimenciones de la base de datos:

```
dim(aduana)
```

```
## [1] 133493     30
```

Vamos a ver de que tipo son los campos de la base de datos:

```
sapply(aduana, class)
```

```
##          TIPO_IMPORTACION       ESTADO_DECLARACION              FEC_INGRESO
##                "character"              "character"                   "Date"
##                  DISTRITO                 REGIMEN                SUBPARTIDA
##                "character"              "character"                "numeric"
##     CODIGO_COMPLEMENTARIO    CODIGO_SUPLEMENTARIO  DESCRIPCION_ARANCELARIA
##                 "numeric"                "numeric"                "character"
##               PAIS_ORIGEN       CODIGO_LIBERACION TRATAMIENTO_PREFERENCIAL
##                "character"              "character"                "character"
##    CONVENIO_INTERNACIONAL      TIPO_UNIDAD_FISICA    TIPO_UNIDAD_COMERCIAL
##                "character"              "character"                "character"
##                  PESO_NETO                      FOB                      FLETE
##                 "numeric"                "numeric"                 "numeric"
##                    SEGURO                      CIF           CANTIDAD_FISICA
##                 "numeric"                "numeric"                 "numeric"
##        CANTIDAD_COMERCIAL             SALVAGUARDIA  SALVAGUARDIA_ESPECIFICA
##                 "numeric"                "numeric"                 "numeric"
##                   FODINFA            ICE_ADVALOREM           ICE_ESPECIFICO
##                 "numeric"                "numeric"                 "numeric"
##                        IVA                ADVALOREM     ADVALOREM_ESPECIFICO
##                 "numeric"                "numeric"                 "numeric"
```

Vamos a revisar si existen elementos en blanco o elementos duplicados:

```
anyNA(aduana)
```

```
## [1] FALSE
```

```
anyDuplicated(aduana)
```

```
## [1] 0
```

Podemos ver que no hay celdas en blanco ni duplicadas.

La base de datos tiene algunas columnas que no son relevantes para nuestro estudio, por lo que vamos a crear una nueva base solo con las columnas que por ahora pensamos que son relevantes y lo vamos a almacenar en una base de datos que se llame **base_seleccionada**:

```
# Definimos las variables de entrada
vars_entrada <- c(
  "REGIMEN", "SUBPARTIDA", "PAIS_ORIGEN", "CONVENIO_INTERNACIONAL",
  "PESO_NETO", "CANTIDAD_FISICA", "CANTIDAD_COMERCIAL",
  "TIPO_UNIDAD_FISICA", "TIPO_UNIDAD_COMERCIAL",
  "FOB", "FLETE", "SEGURO", "CIF"
)

# Definimos las variables de salida
vars_salida <- c(
  "ADVALOREM","ICE_ADVALOREM",
  "FODINFA", "IVA"
```

```
)

# Unimos las variables seleccionadas
vars_seleccionadas <- c(vars_entrada, vars_salida)

# Creamos la nueva base de datos
base_seleccionada <- aduana[, vars_seleccionadas]

# Verificación básica
dim(base_seleccionada)
```

```
## [1] 133493     17
```

Analisis multivariante

Para saber si nuestro proyecto es Viable, voy a obtener la matriz de acrianzaras, para analizar si las variables están relacionadas y podría predecir unas en función de otras.

Sin embargo, previo a esto voy a reemplazar las filas que tienen texto por números, y voy a almacenar los resultados en una variable llamada **base_numerica**

```
# Se identificaron automáticamente las columnas que son de tipo carácter (texto)
vars_char <- names(base_seleccionada)[sapply(base_seleccionada, is.character)]

# Se creó una copia de la base original para trabajar con la versión numérica
base_numerica <- base_seleccionada

# Se transformaron las variables categóricas a valores numéricos mediante codificación de factores
base_numerica[vars_char] <- lapply(base_numerica[vars_char], function(x) {
  as.numeric(as.factor(x))
})

# Se verificó el tipo de dato de todas las columnas luego de la transformación
sapply(base_numerica, class)
```

```
##                REGIMEN             SUBPARTIDA            PAIS_ORIGEN
##              "numeric"              "numeric"              "numeric"
## CONVENIO_INTERNACIONAL             PESO_NETO        CANTIDAD_FISICA
##              "numeric"              "numeric"              "numeric"
##     CANTIDAD_COMERCIAL     TIPO_UNIDAD_FISICA  TIPO_UNIDAD_COMERCIAL
##              "numeric"              "numeric"              "numeric"
##                    FOB                  FLETE                 SEGURO
##              "numeric"              "numeric"              "numeric"
##                    CIF              ADVALOREM          ICE_ADVALOREM
##              "numeric"              "numeric"              "numeric"
##                FODINFA                    IVA
##              "numeric"              "numeric"
```

Nota: Yo solía hacer esto a mano, investigando en Internet, vi que había dos formas automáticas de hacerlo, la primera usando el método de factores y la segunda usando el método "One-Hot Encoding", la primera sirve para modelos basados en árboles y métodos de clasificación que no dependen de relaciones lineales, como voy a usar random forest en mi proyecto decidí usar esa para no hacer mas grande la base.

Ahora voy a normalizar la base y la voy a guardar en una nueva base de datos llamada **base_normalizada**

```r
# Función de normalización Min-Max
normalizar_minmax <- function(x) {
  (x - min(x)) / (max(x) - min(x))
}

# Aplicación de la normalización a todas las columnas
base_normalizada <- as.data.frame(
  lapply(base_numerica, normalizar_minmax)
)

# Verificación
summary(base_normalizada)
```

```
##     REGIMEN         SUBPARTIDA      PAIS_ORIGEN     CONVENIO_INTERNACIONAL
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.4866   1st Qu.:0.1712   1st Qu.:1.0000
##  Median :0.0000   Median :0.8448   Median :0.2123   Median :1.0000
##  Mean   :0.0262   Mean   :0.6964   Mean   :0.4007   Mean   :0.9304
##  3rd Qu.:0.0000   3rd Qu.:0.8666   3rd Qu.:0.6233   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##    PESO_NETO        CANTIDAD_FISICA    CANTIDAD_COMERCIAL  TIPO_UNIDAD_FISICA
##  Min.   :0.00e+00   Min.   :0.00e+00   Min.   :0.0000000   Min.   :0.0000
##  1st Qu.:0.00e+00   1st Qu.:0.00e+00   1st Qu.:0.0000003   1st Qu.:0.0000
##  Median :4.00e-07   Median :1.00e-07   Median :0.0000027   Median :0.0000
##  Mean   :1.69e-04   Mean   :3.95e-05   Mean   :0.0006936   Mean   :0.1189
##  3rd Qu.:4.50e-06   3rd Qu.:1.00e-06   3rd Qu.:0.0000385   3rd Qu.:0.3750
##  Max.   :1.00e+00   Max.   :1.00e+00   Max.   :1.0000000   Max.   :1.0000
##  TIPO_UNIDAD_COMERCIAL      FOB                FLETE
##  Min.   :0.0000        Min.   :0.0000000   Min.   :0.0000000
##  1st Qu.:0.7761        1st Qu.:0.0000011   1st Qu.:0.0000000
##  Median :0.9254        Median :0.0000060   Median :0.0000031
##  Mean   :0.7926        Mean   :0.0001532   Mean   :0.0002327
##  3rd Qu.:0.9254        3rd Qu.:0.0000392   3rd Qu.:0.0000412
##  Max.   :1.0000        Max.   :1.0000000   Max.   :1.0000000
##      SEGURO              CIF               ADVALOREM         ICE_ADVALOREM
##  Min.   :0.00e+00   Min.   :0.0000000   Min.   :0.0000000   Min.   :0.0000000
##  1st Qu.:2.00e-07   1st Qu.:0.0000011   1st Qu.:0.0000000   1st Qu.:0.0000000
##  Median :1.60e-06   Median :0.0000063   Median :0.0000000   Median :0.0000000
##  Mean   :7.79e-05   Mean   :0.0001549   Mean   :0.0006829   Mean   :0.0001248
##  3rd Qu.:1.23e-05   3rd Qu.:0.0000406   3rd Qu.:0.0000458   3rd Qu.:0.0000000
##  Max.   :1.00e+00   Max.   :1.0000000   Max.   :1.0000000   Max.   :1.0000000
##     FODINFA               IVA
##  Min.   :0.0000000   Min.   :0.0000000
##  1st Qu.:0.0000170   1st Qu.:0.0000131
##  Median :0.0000933   Median :0.0000774
##  Mean   :0.0017479   Mean   :0.0013989
##  3rd Qu.:0.0006068   3rd Qu.:0.0004856
##  Max.   :1.0000000   Max.   :1.0000000
```

Ahora voy a calcular la matriz de covarianzas a partir de la base de datos normalizada, con el fin de analizar el grado de variación conjunta entre las variables de entrada y salida, y evaluar su relación estadística previa a la construcción del modelo.

```r
# Se calculó la matriz de covarianzas de la base de datos normalizada
matriz_covarianzas <- cov(base_normalizada)

# Se visualiza la matriz de covarianzas
matriz_covarianzas
```

```
##                            REGIMEN    SUBPARTIDA   PAIS_ORIGEN
## REGIMEN                 1.616696e-02 -1.561619e-03  1.926938e-03
## SUBPARTIDA             -1.561619e-03  5.263111e-02 -1.135358e-04
## PAIS_ORIGEN             1.926938e-03 -1.135358e-04  9.417821e-02
## CONVENIO_INTERNACIONAL  1.443411e-03  9.003622e-03  1.006703e-02
## PESO_NETO              -1.474008e-06 -6.023665e-05  3.364288e-05
## CANTIDAD_FISICA        -1.028605e-07 -1.077666e-05  1.838823e-06
## CANTIDAD_COMERCIAL      1.084588e-05 -1.030239e-04 -3.990288e-05
## TIPO_UNIDAD_FISICA      9.194656e-04 -2.699766e-02 -1.470648e-03
## TIPO_UNIDAD_COMERCIAL   8.988485e-04  1.474206e-02  4.177749e-03
## FOB                     1.947331e-06 -2.606900e-05  1.507368e-05
## FLETE                   4.304630e-06 -3.922680e-06 -2.416972e-05
## SEGURO                  8.193631e-07 -1.650565e-05  1.713149e-05
## CIF                     1.982161e-06 -2.601623e-05  1.483578e-05
## ADVALOREM               3.900552e-05  2.060546e-08 -2.615045e-05
## ICE_ADVALOREM           2.945079e-05  3.492221e-06 -5.771044e-06
## FODINFA                 2.957488e-05 -2.203677e-04 -4.006988e-05
## IVA                     4.338757e-05 -3.569464e-05 -4.441851e-05
##                        CONVENIO_INTERNACIONAL     PESO_NETO CANTIDAD_FISICA
## REGIMEN                         1.443411e-03 -1.474008e-06   -1.028605e-07
## SUBPARTIDA                      9.003622e-03 -6.023665e-05   -1.077666e-05
## PAIS_ORIGEN                     1.006703e-02  3.364288e-05    1.838823e-06
## CONVENIO_INTERNACIONAL          3.753996e-02 -2.963903e-06   -6.559986e-06
## PESO_NETO                      -2.963903e-06  3.241153e-05    1.732109e-06
## CANTIDAD_FISICA                -6.559986e-06  1.732109e-06    8.103111e-06
## CANTIDAD_COMERCIAL             -3.322426e-05  2.646483e-06    1.725922e-06
## TIPO_UNIDAD_FISICA             -7.501900e-03  6.297668e-05    4.418081e-06
## TIPO_UNIDAD_COMERCIAL           5.290064e-03 -4.339660e-05   -5.409738e-06
## FOB                            -4.699734e-06  1.502080e-05    5.867852e-07
## FLETE                          -1.580879e-05  1.066614e-06    1.697545e-07
## SEGURO                          1.829995e-06  1.438033e-05    4.467142e-07
## CIF                            -4.804013e-06  1.502586e-05    5.872124e-07
## ADVALOREM                      -1.352746e-05  4.065337e-07    3.544244e-08
## ICE_ADVALOREM                  -1.999027e-05  6.585341e-08   -4.267774e-09
## FODINFA                        -1.234064e-04  1.748871e-05    3.194617e-06
## IVA                            -7.065090e-05  3.831000e-06    7.620509e-07
##                        CANTIDAD_COMERCIAL TIPO_UNIDAD_FISICA
## REGIMEN                      1.084588e-05       9.194656e-04
## SUBPARTIDA                  -1.030239e-04      -2.699766e-02
## PAIS_ORIGEN                 -3.990288e-05      -1.470648e-03
## CONVENIO_INTERNACIONAL      -3.322426e-05      -7.501900e-03
## PESO_NETO                    2.646483e-06       6.297668e-05
## CANTIDAD_FISICA              1.725922e-06       4.418081e-06
## CANTIDAD_COMERCIAL           9.753945e-05       8.708331e-05
## TIPO_UNIDAD_FISICA           8.708331e-05       3.988078e-02
## TIPO_UNIDAD_COMERCIAL       -6.665800e-05      -1.445844e-02
## FOB                          1.524595e-06       2.924587e-05
```

```
## FLETE                       9.983546e-07       1.247655e-05
## SEGURO                      9.863549e-07       2.094649e-05
## CIF                         1.529934e-06       2.929702e-05
## ADVALOREM                   1.187676e-06       1.532162e-05
## ICE_ADVALOREM              -5.112984e-08       3.088963e-06
## FODINFA                     1.308173e-05       1.898436e-04
## IVA                         8.375499e-06       9.544370e-05
##                       TIPO_UNIDAD_COMERCIAL           FOB         FLETE
## REGIMEN                        8.988485e-04  1.947331e-06  4.304630e-06
## SUBPARTIDA                     1.474206e-02 -2.606900e-05 -3.922680e-06
## PAIS_ORIGEN                    4.177749e-03  1.507368e-05 -2.416972e-05
## CONVENIO_INTERNACIONAL         5.290064e-03 -4.699734e-06 -1.580879e-05
## PESO_NETO                     -4.339660e-05  1.502080e-05  1.066614e-06
## CANTIDAD_FISICA               -5.409738e-06  5.867852e-07  1.697545e-07
## CANTIDAD_COMERCIAL            -6.665800e-05  1.524595e-06  9.983546e-07
## TIPO_UNIDAD_FISICA            -1.445844e-02  2.924587e-05  1.247655e-05
## TIPO_UNIDAD_COMERCIAL          4.543034e-02 -2.396980e-05 -6.346723e-06
## FOB                           -2.396980e-05  1.500011e-05  7.669999e-07
## FLETE                         -6.346723e-06  7.669999e-07  1.452090e-05
## SEGURO                        -2.016140e-05  1.471164e-05  2.815717e-07
## CIF                           -2.399991e-05  1.500545e-05  9.173613e-07
## ADVALOREM                     -7.876909e-06  1.908966e-06  6.843284e-06
## ICE_ADVALOREM                  1.535875e-05  5.718672e-07  1.710295e-06
## FODINFA                       -9.387516e-05  8.489228e-06  1.413289e-05
## IVA                           -5.087482e-05  4.747096e-06  1.125904e-05
##                                     SEGURO           CIF     ADVALOREM ICE_ADVALOREM
## REGIMEN                       8.193631e-07  1.982161e-06  3.900552e-05  2.945079e-05
## SUBPARTIDA                   -1.650565e-05 -2.601623e-05  2.060546e-08  3.492221e-06
## PAIS_ORIGEN                   1.713149e-05  1.483578e-05 -2.615045e-05 -5.771044e-06
## CONVENIO_INTERNACIONAL        1.829995e-06 -4.804013e-06 -1.352746e-05 -1.999027e-05
## PESO_NETO                     1.438033e-05  1.502586e-05  4.065337e-07  6.585341e-08
## CANTIDAD_FISICA               4.467142e-07  5.872124e-07  3.544244e-08 -4.267774e-09
## CANTIDAD_COMERCIAL            9.863549e-07  1.529934e-06  1.187676e-06 -5.112984e-08
## TIPO_UNIDAD_FISICA            2.094649e-05  2.929702e-05  1.532162e-05  3.088963e-06
## TIPO_UNIDAD_COMERCIAL        -2.016140e-05 -2.399991e-05 -7.876909e-06  1.535875e-05
## FOB                           1.471164e-05  1.500545e-05  1.908966e-06  5.718672e-07
## FLETE                         2.815717e-07  9.173613e-07  6.843284e-06  1.710295e-06
## SEGURO                        1.480107e-05  1.471554e-05  2.743040e-07  5.317923e-08
## CIF                           1.471554e-05  1.501238e-05  1.965907e-06  5.850076e-07
## ADVALOREM                     2.743040e-07  1.965907e-06  7.940241e-05  2.420984e-05
## ICE_ADVALOREM                 5.317923e-08  5.850076e-07  2.420984e-05  1.758665e-05
## FODINFA                       2.058020e-06  8.576575e-06  3.194103e-05  9.433492e-06
## IVA                           1.336445e-06  4.833639e-06  4.200632e-05  1.450185e-05
##                                   FODINFA           IVA
## REGIMEN                       2.957488e-05  4.338757e-05
## SUBPARTIDA                   -2.203677e-04 -3.569464e-05
## PAIS_ORIGEN                  -4.006988e-05 -4.441851e-05
## CONVENIO_INTERNACIONAL       -1.234064e-04 -7.065090e-05
## PESO_NETO                     1.748871e-05  3.831000e-06
## CANTIDAD_FISICA               3.194617e-06  7.620509e-07
## CANTIDAD_COMERCIAL            1.308173e-05  8.375499e-06
## TIPO_UNIDAD_FISICA            1.898436e-04  9.544370e-05
## TIPO_UNIDAD_COMERCIAL        -9.387516e-05 -5.087482e-05
## FOB                           8.489228e-06  4.747096e-06
```

```
## FLETE                   1.413289e-05  1.125904e-05
## SEGURO                  2.058020e-06  1.336445e-06
## CIF                     8.576575e-06  4.833639e-06
## ADVALOREM               3.194103e-05  4.200632e-05
## ICE_ADVALOREM           9.433492e-06  1.450185e-05
## FODINFA                 1.414210e-04  7.809524e-05
## IVA                     7.809524e-05  8.577755e-05
```

Esta matriz, me suele parecer un poco confusa, por lo que personalmente uso la matriz de correlaciones:

```
# Se calculó la matriz de covarianzas de la base de datos normalizada
matriz_correlacion <- cor(base_normalizada)

# Se visualiza la matriz de covarianzas
matriz_correlacion
```

```
##                            REGIMEN    SUBPARTIDA  PAIS_ORIGEN
## REGIMEN                 1.0000000000 -5.353521e-02  0.049383101
## SUBPARTIDA             -0.0535352133  1.000000e+00 -0.001612636
## PAIS_ORIGEN             0.0493831008 -1.612636e-03  1.000000000
## CONVENIO_INTERNACIONAL  0.0585907313  2.025578e-01  0.169308720
## PESO_NETO              -0.0020362720 -4.612005e-02  0.019256090
## CANTIDAD_FISICA        -0.0002841899 -1.650201e-02  0.002104936
## CANTIDAD_COMERCIAL      0.0086369508 -4.547019e-02 -0.013165546
## TIPO_UNIDAD_FISICA      0.0362109059 -5.892818e-01 -0.023996713
## TIPO_UNIDAD_COMERCIAL   0.0331664716  3.014837e-01  0.063869592
## FOB                     0.0039543788 -2.933972e-02  0.012682262
## FLETE                   0.0088843343 -4.487091e-03 -0.020668057
## SEGURO                  0.0016750031 -1.870100e-02  0.014510201
## CIF                     0.0040234623 -2.926836e-02  0.012477003
## ADVALOREM               0.0344266645  1.007962e-05 -0.009562844
## ICE_ADVALOREM           0.0552320468  3.629852e-03 -0.004484227
## FODINFA                 0.0195592323 -8.077365e-02 -0.010979583
## IVA                     0.0368437905 -1.679945e-02 -0.015627949
##                        CONVENIO_INTERNACIONAL    PESO_NETO CANTIDAD_FISICA
## REGIMEN                           0.058590731 -0.002036272   -0.0002841899
## SUBPARTIDA                        0.202557841 -0.046120047   -0.0165020127
## PAIS_ORIGEN                       0.169308720  0.019256090    0.0021049361
## CONVENIO_INTERNACIONAL            1.000000000 -0.002686998   -0.0118940595
## PESO_NETO                        -0.002686998  1.000000000    0.1068807819
## CANTIDAD_FISICA                  -0.011894059  0.106880782    1.0000000000
## CANTIDAD_COMERCIAL               -0.017362738  0.047068375    0.0613910549
## TIPO_UNIDAD_FISICA               -0.193884226  0.055392135    0.0077718777
## TIPO_UNIDAD_COMERCIAL             0.128097601 -0.035762916   -0.0089161501
## FOB                              -0.006262953  0.681233731    0.0532238227
## FLETE                            -0.021411892  0.049165513    0.0156494364
## SEGURO                            0.002455027  0.656556984    0.0407903447
## CIF                              -0.006399300  0.681184451    0.0532407898
## ADVALOREM                        -0.007835244  0.008013642    0.0013972716
## ICE_ADVALOREM                    -0.024602556  0.002758271   -0.0003575065
## FODINFA                          -0.053559164  0.258315762    0.0943704325
## IVA                              -0.039371679  0.072656674    0.0289048862
##                       CANTIDAD_COMERCIAL TIPO_UNIDAD_FISICA
```

```
## REGIMEN                       0.008636951        0.036210906
## SUBPARTIDA                   -0.045470189       -0.589281839
## PAIS_ORIGEN                  -0.013165546       -0.023996713
## CONVENIO_INTERNACIONAL       -0.017362738       -0.193884226
## PESO_NETO                     0.047068375        0.055392135
## CANTIDAD_FISICA               0.061391055        0.007771878
## CANTIDAD_COMERCIAL            1.000000000        0.044153278
## TIPO_UNIDAD_FISICA            0.044153278        1.000000000
## TIPO_UNIDAD_COMERCIAL        -0.031665704       -0.339677296
## FOB                           0.039858138        0.037812505
## FLETE                         0.026527599        0.016395166
## SEGURO                        0.025959512        0.027263575
## CIF                           0.039981369        0.037863152
## ADVALOREM                     0.013495561        0.008610057
## ICE_ADVALOREM                -0.001234505        0.003688411
## FODINFA                       0.111382763        0.079938719
## IVA                           0.091565892        0.051603432
##                        TIPO_UNIDAD_COMERCIAL          FOB        FLETE
## REGIMEN                            0.033166472  0.003954379  0.008884334
## SUBPARTIDA                         0.301483732 -0.029339718 -0.004487091
## PAIS_ORIGEN                        0.063869592  0.012682262 -0.020668057
## CONVENIO_INTERNACIONAL             0.128097601 -0.006262953 -0.021411892
## PESO_NETO                         -0.035762916  0.681233731  0.049165513
## CANTIDAD_FISICA                   -0.008916150  0.053223823  0.015649436
## CANTIDAD_COMERCIAL                -0.031665704  0.039858138  0.026527599
## TIPO_UNIDAD_FISICA                -0.339677296  0.037812505  0.016395166
## TIPO_UNIDAD_COMERCIAL              1.000000000 -0.029036495 -0.007814115
## FOB                               -0.029036495  1.000000000  0.051969834
## FLETE                             -0.007814115  0.051969834  1.000000000
## SEGURO                            -0.024586746  0.987341511  0.019206391
## CIF                               -0.029061088  0.999947025  0.062132504
## ADVALOREM                         -0.004147305  0.055313877  0.201535243
## ICE_ADVALOREM                      0.017182690  0.035209217  0.107024384
## FODINFA                           -0.037035711  0.184316403  0.311872640
## IVA                               -0.025771701  0.132340861  0.319020013
##                             SEGURO          CIF     ADVALOREM ICE_ADVALOREM
## REGIMEN                 0.001675003  0.004023462  3.442666e-02  0.0552320468
## SUBPARTIDA             -0.018700997 -0.029268357  1.007962e-05  0.0036298519
## PAIS_ORIGEN             0.014510201  0.012477003 -9.562844e-03 -0.0044842267
## CONVENIO_INTERNACIONAL  0.002455027 -0.006399300 -7.835244e-03 -0.0246025557
## PESO_NETO               0.656556984  0.681184451  8.013642e-03  0.0027582707
## CANTIDAD_FISICA         0.040790345  0.053240790  1.397272e-03 -0.0003575065
## CANTIDAD_COMERCIAL      0.025959512  0.039981369  1.349556e-02 -0.0012345045
## TIPO_UNIDAD_FISICA      0.027263575  0.037863152  8.610057e-03  0.0036884109
## TIPO_UNIDAD_COMERCIAL  -0.024586746 -0.029061088 -4.147305e-03  0.0171826905
## FOB                     0.987341511  0.999947025  5.531388e-02  0.0352092169
## FLETE                   0.019206391  0.062132504  2.015352e-01  0.1070243842
## SEGURO                  1.000000000  0.987199113  8.001452e-03  0.0032961269
## CIF                     0.987199113  1.000000000  5.694051e-02  0.0360035298
## ADVALOREM               0.008001452  0.056940510  1.000000e+00  0.6478633267
## ICE_ADVALOREM           0.003296127  0.036003530  6.478633e-01  1.0000000000
## FODINFA                 0.044982760  0.186136717  3.014221e-01  0.1891576263
## IVA                     0.037507469  0.134698447  5.089921e-01  0.3733748165
##                            FODINFA         IVA
```

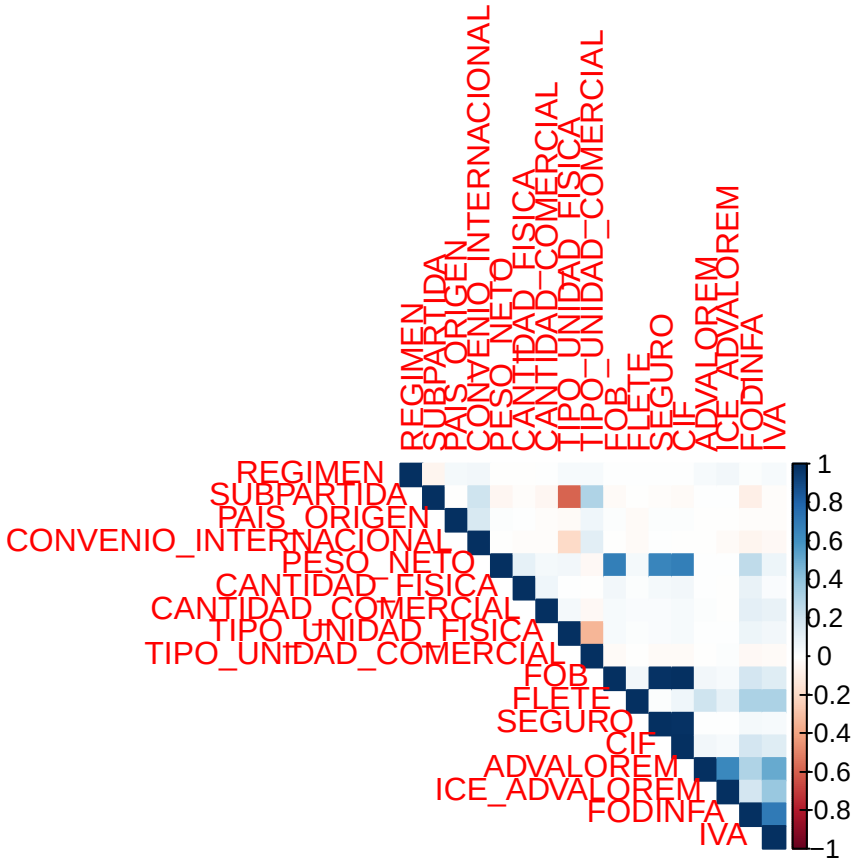
```
## REGIMEN                 0.01955923  0.03684379
## SUBPARTIDA             -0.08077365 -0.01679945
## PAIS_ORIGEN            -0.01097958 -0.01562795
## CONVENIO_INTERNACIONAL -0.05355916 -0.03937168
## PESO_NETO               0.25831576  0.07265667
## CANTIDAD_FISICA         0.09437043  0.02890489
## CANTIDAD_COMERCIAL      0.11138276  0.09156589
## TIPO_UNIDAD_FISICA      0.07993872  0.05160343
## TIPO_UNIDAD_COMERCIAL  -0.03703571 -0.02577170
## FOB                     0.18431640  0.13234086
## FLETE                   0.31187264  0.31902001
## SEGURO                  0.04498276  0.03750747
## CIF                     0.18613672  0.13469845
## ADVALOREM               0.30142211  0.50899207
## ICE_ADVALOREM           0.18915763  0.37337482
## FODINFA                 1.00000000  0.70905631
## IVA                     0.70905631  1.00000000
```

Voy a representar la matriz de correlaciones usando un mapa de calor, para analizar de forma visual la intensidad y el signo de la relación entre las variables de entrada y salida. Este gráfico facilitó la identificación de dependencias fuertes, débiles y relaciones inversas entre variables.

```
library(corrplot)
```

```
## corrplot 0.95 loaded
```

```
# Se generó el gráfico de correlación
corrplot(matriz_correlacion, method = "color", type = "upper")
```



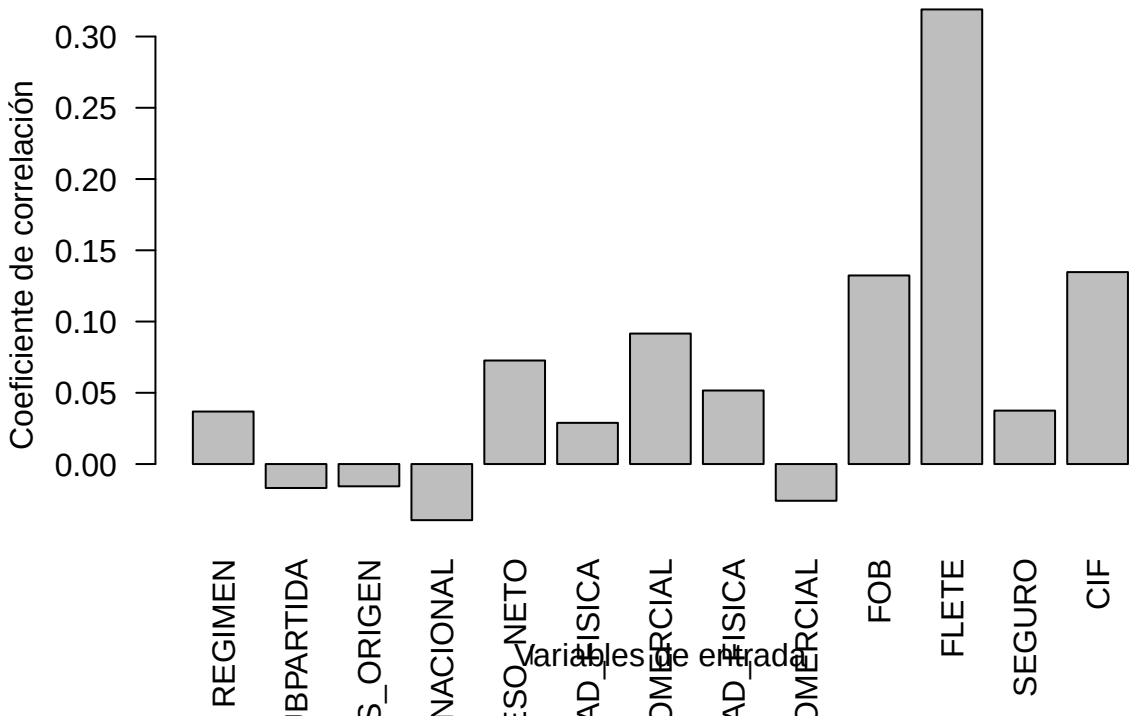
Nota: Tuve que instalar la librería "corrplot"

Para ver los resultados por separado, voy a generar gráficos de barras para cada una de las salidas, quiero ver cuales de las variables de entrada tienen mayor impacto en las salidas.

Esto lo hago por que no un gran numero de variables de entrada y no quiero usarlas todas, solo las 4 mas relevantes.

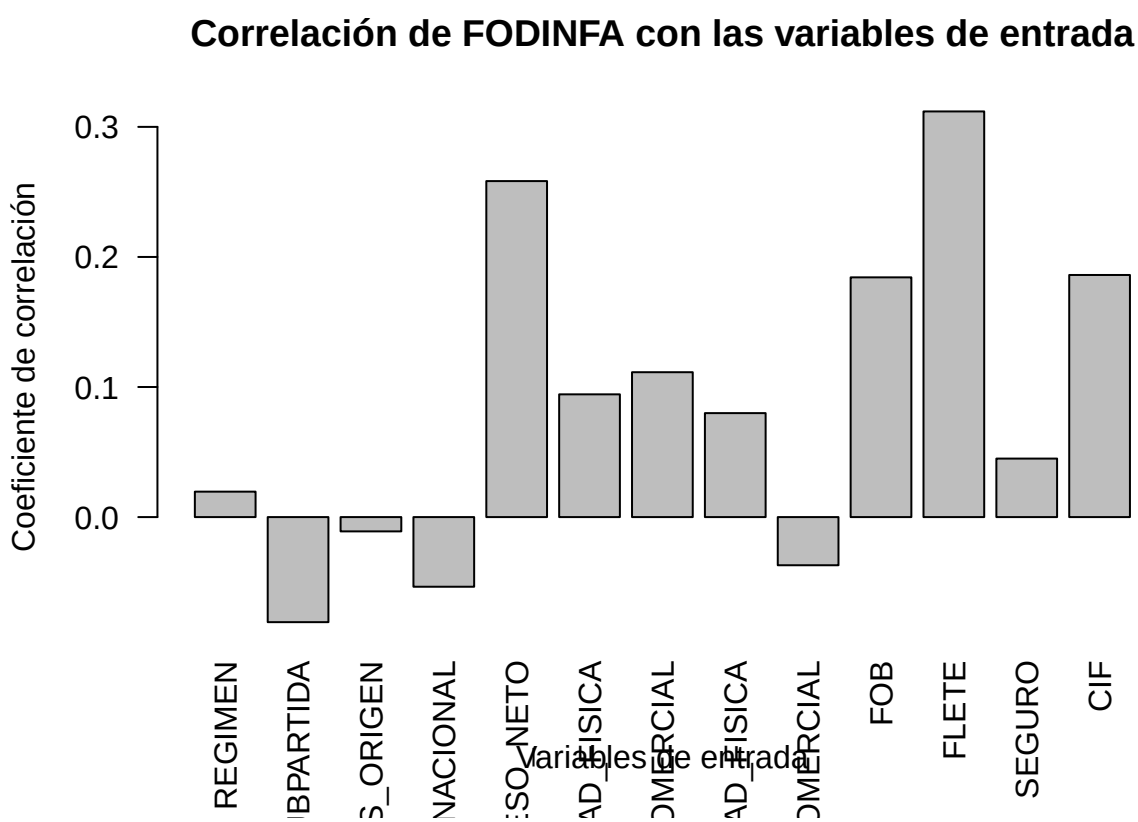
Para el IVA:

```
# Se extrajo la correlación de IVA con todas las variables de entrada
cor_iva <- matriz_correlacion[vars_entrada, "IVA"]

# Se generó el gráfico de barras de las correlaciones
barplot(cor_iva,
        las = 2,                  # rota etiquetas del eje X
        main = "Correlación de IVA con las variables de entrada",
        ylab = "Coeficiente de correlación",
        xlab = "Variables de entrada")
```



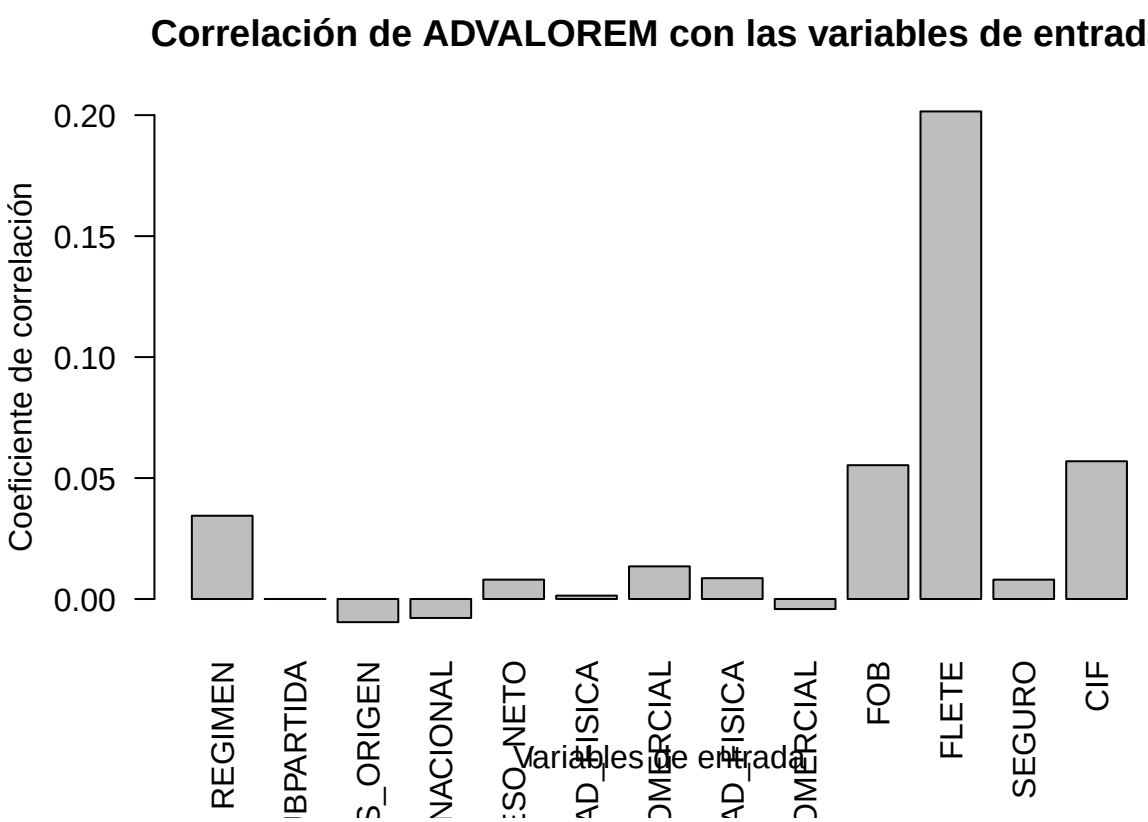
Para el FODINFA:

```r
cor_FODINFA <- matriz_correlacion[vars_entrada, "FODINFA"]

# Se generó el gráfico de barras de las correlaciones
barplot(cor_FODINFA,
        las = 2,                 # rota etiquetas del eje X
        main = "Correlación de FODINFA con las variables de entrada",
        ylab = "Coeficiente de correlación",
        xlab = "Variables de entrada")
```



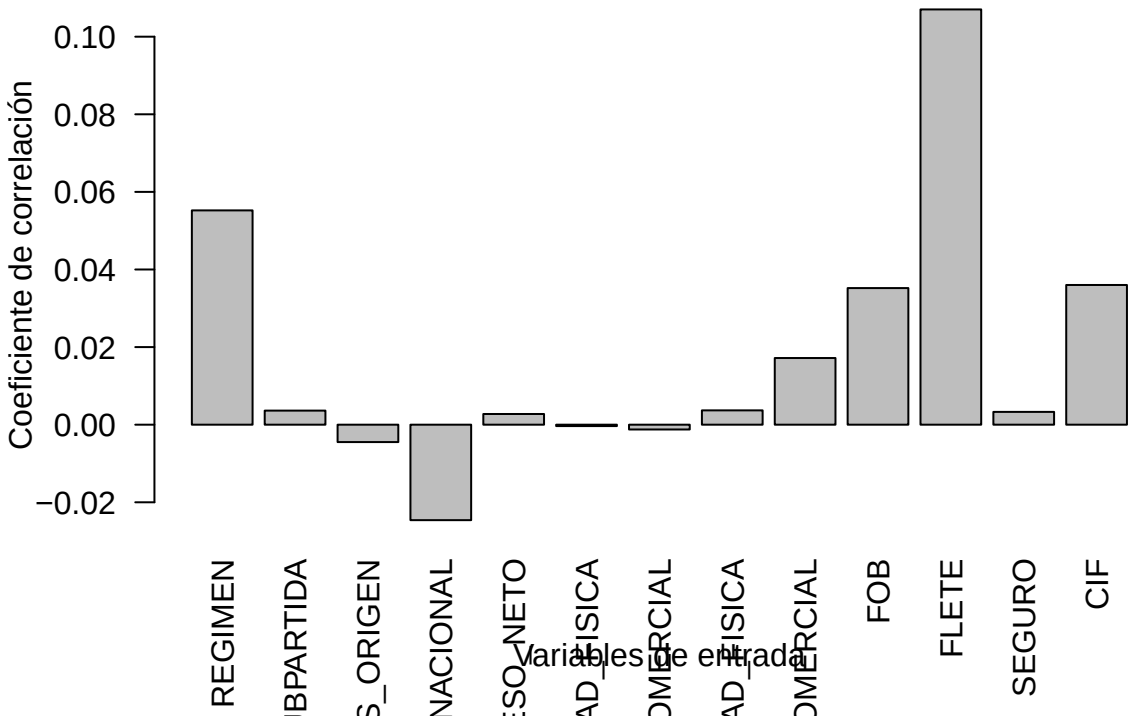
Para el ADVALOREM:

```
cor_ADVALOREM <- matriz_correlacion[vars_entrada, "ADVALOREM"]

# Se generó el gráfico de barras de las correlaciones
barplot(cor_ADVALOREM,
        las = 2,                 # rota etiquetas del eje X
        main = "Correlación de ADVALOREM con las variables de entrada",
        ylab = "Coeficiente de correlación",
        xlab = "Variables de entrada")
```



Para el ICE_ADVALOREM:

```
cor_ICE_ADVALOREM <- matriz_correlacion[vars_entrada, "ICE_ADVALOREM"]

# Se generó el gráfico de barras de las correlaciones
barplot(cor_ICE_ADVALOREM,
        las = 2,                  # rota etiquetas del eje X
        main = "Correlación de ICE_ADVALOREM con las variables de entrada",
        ylab = "Coeficiente de correlación",
        xlab = "Variables de entrada")
```



Resultados

A partir del análisis de correlación se observa que, para la salida **IVA**, las variables de entrada con mayor impacto son:

- `FLETE`
- `FOB`
- `CIF`
- `TIPO_UNIDAD_COMERCIAL`

Para la salida **FODINFA**, las variables de entrada de mayor influencia son:

- `FLETE`
- `PESO_NETO`
- `CIF`
- `FOB`

Finalmente, para las salidas **ADVALOREM** e **ICE_ADVALOREM**, las variables de entrada más relevantes son:

- `FLETE`
- `FOB`
- `CIF`
- `REGIMEN`

De manera general, se puede concluir que las variables de entrada **FLETE, FOB y CIF** presentan el mayor impacto sobre todas las salidas analizadas, mientras que una cuarta variable (como TIPO_UNIDAD_COMERCIAL, PESO_NETO o REGIMEN) aporta información adicional específica dependiendo de cada tributo.

Resultados del Análisis Multivariable

Como resultado del estudio, se analizó el impacto de las variables de entrada sobre las salidas del modelo, identificando aquellas con mayor relevancia para la predicción de los tributos. De esta forma, se pasó de considerar inicialmente **13 variables de entrada** a trabajar únicamente con **4 variables principales**.

En la siguiente etapa se generarán **cuatro conjuntos de datos**, uno para cada variable de salida, en los cuales las tres primeras columnas corresponderán a las variables de entrada seleccionadas y la última columna a la salida específica de cada modelo.

```
# Se crearon las bases de datos específicas para cada tributo
# usando como variables de entrada FLETE, FOB y CIF
# y como variable de salida el tributo correspondiente

base_IVA <- aduana[, c("FLETE", "FOB", "CIF","TIPO_UNIDAD_COMERCIAL", "IVA")]
base_ADVALOREM <- aduana[, c("FLETE", "FOB", "CIF", "ADVALOREM")]
base_ICE_ADVALOREM <- aduana[, c("FLETE", "FOB", "CIF", "ICE_ADVALOREM")]
base_FODINFA <- aduana[, c("FLETE", "FOB", "CIF", "FODINFA")]

# Verificación rápida de las cuatro bases creadas
list(
  IVA = dim(base_IVA),
  ADVALOREM = dim(base_ADVALOREM),
  ICE_ADVALOREM = dim(base_ICE_ADVALOREM),
  FODINFA = dim(base_FODINFA)
)
```

```
## $IVA
## [1] 133493      5
##
## $ADVALOREM
## [1] 133493      4
##
## $ICE_ADVALOREM
## [1] 133493      4
##
## $FODINFA
## [1] 133493      4
```

Modelo Random Forest

Con el fin de modelar la relación no lineal entre las variables de entrada (FLETE, FOB y CIF) y los tributos aduaneros de salida (IVA, ADVALOREM, ICE_ADVALOREM y FODINFA), se emplea el algoritmo **Random Forest de regresión**. Este método basado en ensambles de árboles de decisión permite capturar interacciones complejas entre las variables, reducir la varianza del modelo y mejorar la capacidad de generalización, siendo especialmente adecuado para conjuntos de datos con relaciones no lineales y posible multicolinealidad entre las variables de entrada.

Primero cargamos la ibreria y nos aseguramos de que cada vez que se corra este programa se obtengan los mismos resultados:

```r
# Se cargó la librería necesaria para Random Forest
library(randomForest)
```

```
## randomForest 4.7-1.2
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```r
# Se fijó la semilla para garantizar reproducibilidad
set.seed(123)
```

Mi idea inicial fue utilizar el 75% de los datos para entrenar el modelo y el 25% restante para validarlo. Sin embargo, la base de datos original contiene **133 493** filas y, al aplicar el modelo de Random Forest sobre todo el conjunto, se presentaba un cuello de botella computacional que podía superar los 20 minutos de ejecución.

Por este motivo, se decidió trabajar con una **muestra aleatoria del 10%** de la base original, equivalente a **13 400 filas**, que se considera representativa para los fines de este trabajo. Sobre esta muestra se utiliza el **70% de los datos para el entrenamiento** de los modelos y el **30% restante para la validación**.

```r
n_total <- nrow(base_numerica)
tam_muestra <- min(13400, n_total)
idx_muestra <- sample(seq_len(n_total), size = tam_muestra)
aduana_muestra <- base_numerica[idx_muestra, ]
dim(aduana_muestra)
```

```
## [1] 13400    17
```

A partir de la muestra se construyeron las bases específicas para cada salida, utilizando como variables de entrada `FLETE`, `FOB` y `CIF`. Posteriormente, cada base se dividió en un conjunto de entrenamiento (70%) y otro de prueba (30%):

```r
#Bases específicas por salida, construidas a partir de la muestra
base_IVA <- aduana_muestra[, c("FLETE", "FOB", "CIF","TIPO_UNIDAD_COMERCIAL", "IVA")]
base_ADVALOREM <- aduana_muestra[, c("FLETE", "FOB", "CIF","REGIMEN", "ADVALOREM")]
base_ICE_ADVALOREM <- aduana_muestra[, c("FLETE", "FOB", "CIF","REGIMEN", "ICE_ADVALOREM")]
base_FODINFA <- aduana_muestra[, c("FLETE", "FOB", "CIF","PESO_NETO", "FODINFA")]

#División en entrenamiento (70%) y prueba (30%)

n_m <- nrow(aduana_muestra)
idx_entrenamiento <- sample(seq_len(n_m), size = floor(0.7 * n_m))

train_IVA <- base_IVA[idx_entrenamiento, ]
test_IVA <- base_IVA[-idx_entrenamiento, ]

train_ADVALOREM <- base_ADVALOREM[idx_entrenamiento, ]
test_ADVALOREM <- base_ADVALOREM[-idx_entrenamiento, ]

train_ICE_ADVALOREM <- base_ICE_ADVALOREM[idx_entrenamiento, ]
test_ICE_ADVALOREM <- base_ICE_ADVALOREM[-idx_entrenamiento, ]

train_FODINFA <- base_FODINFA[idx_entrenamiento, ]
test_FODINFA <- base_FODINFA[-idx_entrenamiento, ]
```

A continuación se ajustan los modelos de Random Forest de regresión para cada una de las salidas consideradas en el estudio:

```
#Modelos Random Forest para cada salida

modelo_IVA <- randomForest(
IVA ~ FLETE + FOB + CIF + TIPO_UNIDAD_COMERCIAL,
data = train_IVA,
ntree = 150,
mtry = 2,
importance = TRUE
)

modelo_ADVALOREM <- randomForest(
ADVALOREM ~ FLETE + FOB + CIF +REGIMEN,
data = train_ADVALOREM,
ntree = 150,
mtry = 2,
importance = TRUE
)

modelo_ICE_ADVALOREM <- randomForest(
ICE_ADVALOREM ~ FLETE + FOB + CIF +REGIMEN,
data = train_ICE_ADVALOREM,
ntree = 150,
mtry = 2,
importance = TRUE
)

modelo_FODINFA <- randomForest(
FODINFA ~ FLETE + FOB + CIF + PESO_NETO,
data = train_FODINFA,
ntree = 150,
mtry = 2,
importance = TRUE
)
```

Con los modelos ajustados, se generan las predicciones sobre el conjunto de prueba y se calculan los indicadores de desempeño RMSE y R^2 para cada salida:

```
# Predicciones sobre el conjunto de prueba

pred_IVA           <- predict(modelo_IVA, test_IVA)
pred_ADVALOREM     <- predict(modelo_ADVALOREM, test_ADVALOREM)
pred_ICE_ADVALOREM <- predict(modelo_ICE_ADVALOREM, test_ICE_ADVALOREM)
pred_FODINFA       <- predict(modelo_FODINFA, test_FODINFA)

# Funciones para RMSE y R²

rmse <- function(real, pred) {
sqrt(mean((real - pred)^2))
}

r2 <- function(real, pred) {
```

```r
1 - sum((real - pred)^2) / sum((real - mean(real))^2)
}

# Cálculo de métricas para cada modelo

resultados_modelos <- data.frame(
Salida = c("IVA", "ADVALOREM", "ICE_ADVALOREM", "FODINFA"),
RMSE = c(
rmse(test_IVA$IVA, pred_IVA),
rmse(test_ADVALOREM$ADVALOREM, pred_ADVALOREM),
rmse(test_ICE_ADVALOREM$ICE_ADVALOREM, pred_ICE_ADVALOREM),
rmse(test_FODINFA$FODINFA, pred_FODINFA)
),
R2 = c(
r2(test_IVA$IVA, pred_IVA),
r2(test_ADVALOREM$ADVALOREM, pred_ADVALOREM),
r2(test_ICE_ADVALOREM$ICE_ADVALOREM, pred_ICE_ADVALOREM),
r2(test_FODINFA$FODINFA, pred_FODINFA)
)
)

resultados_modelos
```

```
##          Salida       RMSE         R2
## 1           IVA 9393.91846  0.5195762
## 2     ADVALOREM 9909.79338 -0.4860262
## 3 ICE_ADVALOREM 3627.66566 -0.2783317
## 4       FODINFA   71.77609  0.9856417
```

Los resultados obtenidos con el modelo Random Forest muestran un comportamiento diferenciado entre las variables de salida analizadas. Para el **IVA**, el coeficiente de determinación alcanzó un valor de $R^2 = 0.5196$, lo que indica una capacidad predictiva moderada. En el caso de **ADVALOREM** e **ICE_ADVALOREM**, los valores negativos de R^2 confirman que el modelo no logra capturar adecuadamente la variabilidad de estos tributos, dado que su cálculo depende principalmente de factores normativos y arancelarios. Por su parte, **FODINFA** presenta nuevamente un ajuste sobresaliente con un $R^2 = 0.9856$, evidenciando una relación altamente determinística con las variables de entrada consideradas.

- **IVA**:
  - Se obtuvo un $R^2 = 0.5196$, lo que indica que el modelo explica aproximadamente el 52% de la variabilidad del impuesto.
  - El valor de RMSE = 9393.92 refleja un error de predicción moderado en la escala original del tributo.
  - Este resultado confirma que el IVA puede ser estimado parcialmente a partir de variables monetarias y logísticas, aunque sigue dependiendo de factores adicionales de carácter normativo.
- **ADVALOREM**:
  - Se obtuvo un $R^2 = -0.4860$, lo que indica que el modelo tiene un desempeño peor que el uso del valor promedio como estimador.
  - El RMSE = 9909.79 evidencia errores elevados en la predicción.
  - Esto confirma que el ADVALOREM está dominado por reglas arancelarias asociadas a la subpartida, el régimen y los convenios, las cuales no pueden ser capturadas adecuadamente por el modelo bajo este enfoque.

- **ICE_ADVALOREM**:
  - El $R^2 = -0.2783$ muestra que el modelo no logra identificar un patrón estadístico útil para esta variable.
  - El RMSE = 3627.67 sigue siendo elevado en relación con la magnitud del impuesto.
  - Este comportamiento confirma que el ICE depende de criterios específicos de tipo de mercancía y normativa tributaria, más que de relaciones continuas entre variables monetarias.
- **FODINFA**:
  - Se alcanzó un $R^2 = 0.9856$, lo que representa un ajuste excelente entre los valores reales y los predichos.
  - El RMSE = 71.78 es bajo en relación con la magnitud típica del tributo.
  - Este resultado confirma que el FODINFA mantiene una relación directa, estable y altamente predecible con las variables de entrada consideradas.

```r
plot(test_IVA$IVA, pred_IVA,
     main = "Predicción vs Valor Real del IVA",
     xlab = "IVA Real",
     ylab = "IVA Predicho",
     pch = 16,
     col = rgb(0, 0, 1, 0.4))

# Recta ideal y = x
abline(a = 0, b = 1, col = "red", lwd = 2)
```