# Cervical cancer (Risk Factors)

Se adjunta un data set con pacientes que podrían o no, tener cáncer cervical (positivo en una biopsia).

El experto de negocio nos advirtió que no le importa si algunas dimensiones no aportan información relevante en un modelo ML, él quiere que todas se incluyan.

El objetivo del ejercicio es poder obtener un modelo consistente que pueda predecir/diagnosticar el resultado de una biopsia dada la información de algún paciente.

Para más información del dataset, puede consultar el siguiente link:
https://archive.ics.uci.edu/dataset/383/cervical+cancer+risk+factors

## Proceso.

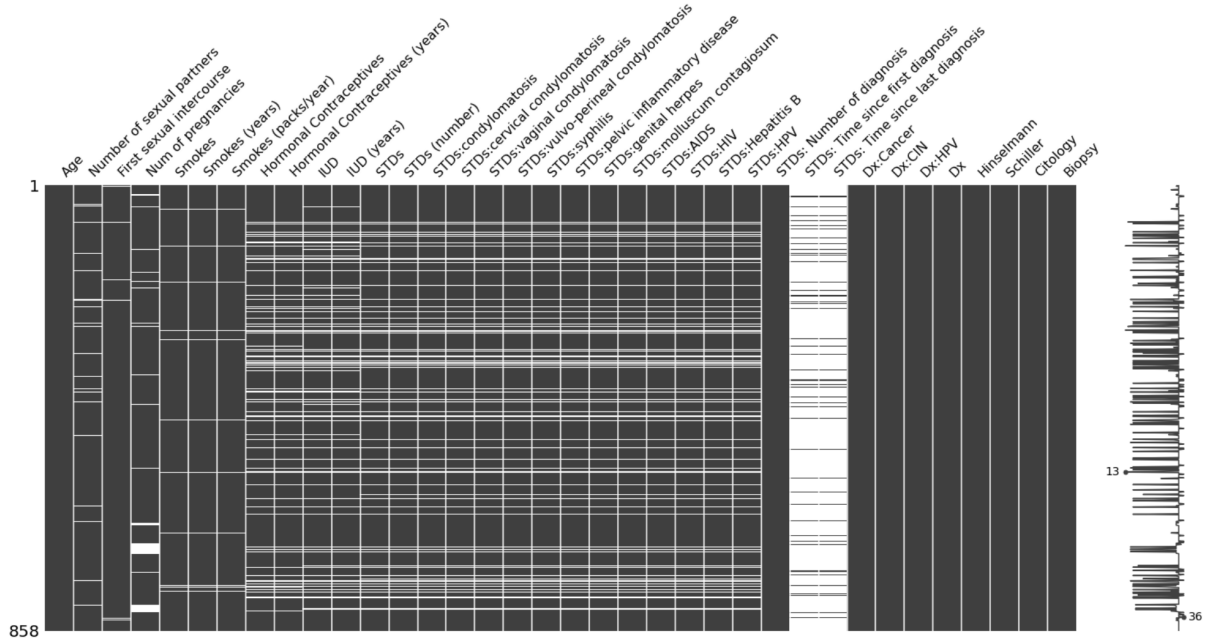
1. Carga de datos y revisión de datos perdidos

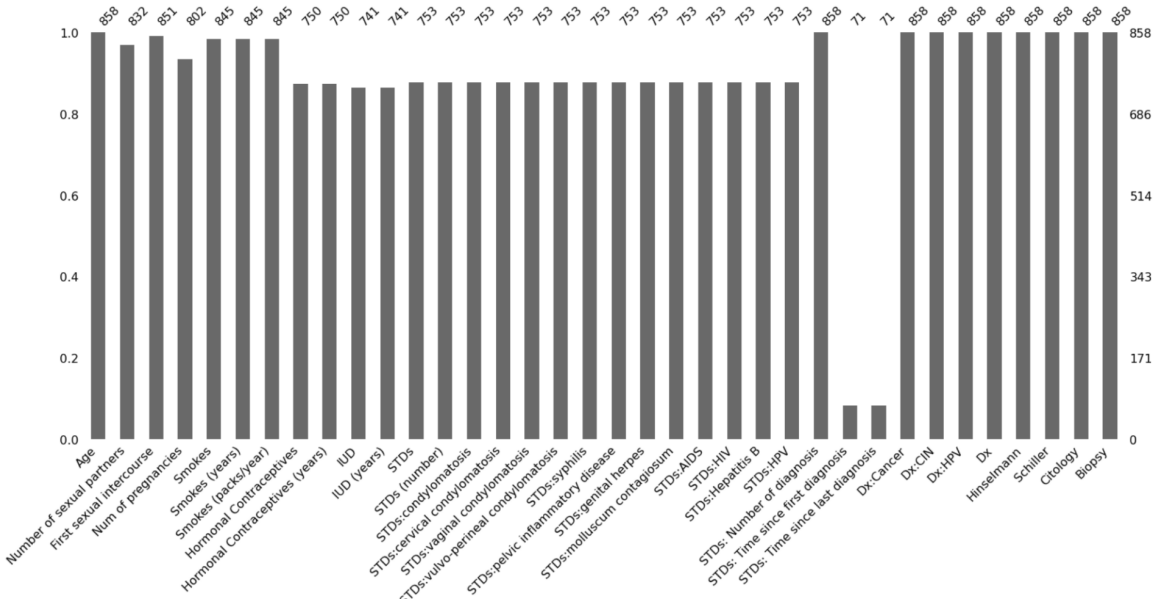
En varias columnas los valores perdidos estaban representados con ? , se reemplazó por el valor NaN

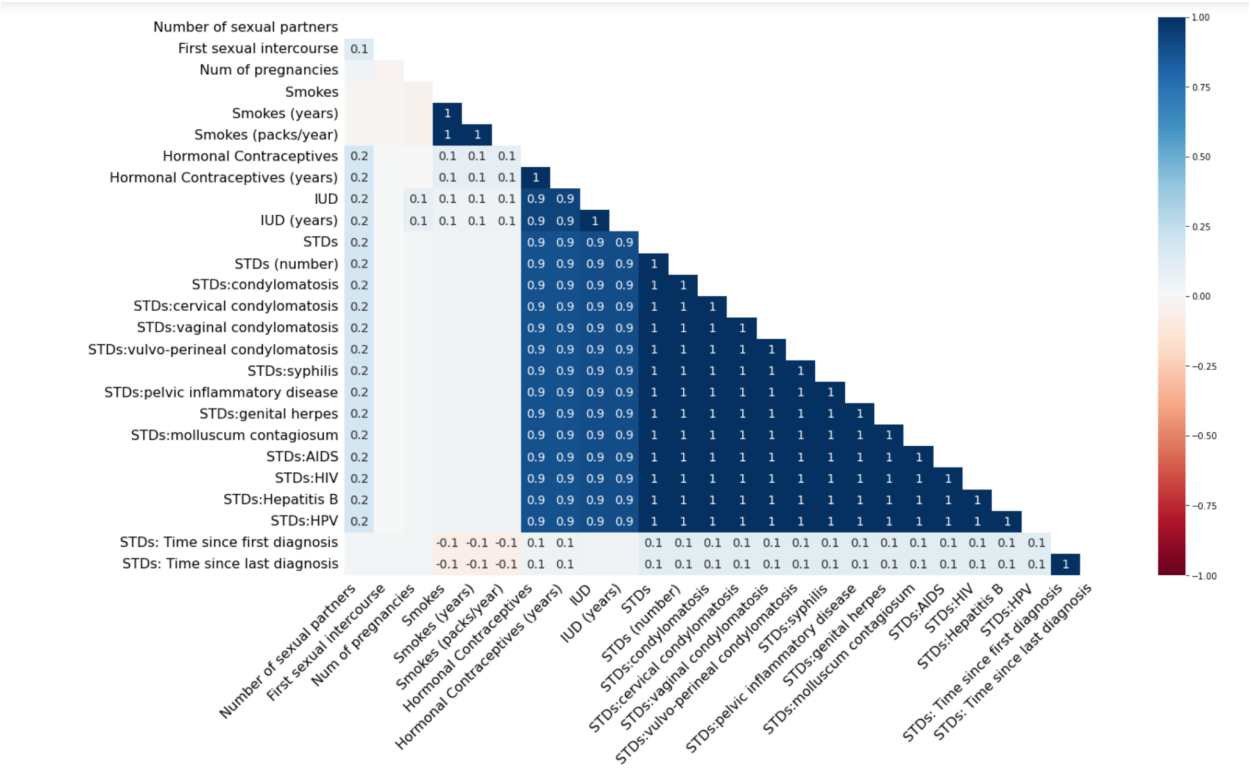
Out[52]:

| | Age | Number of sexual partners | First sexual intercourse | Num of pregnancies | Smokes | Smokes (years) | Smokes (packs/year) | Hormonal Contraceptives | Hormonal Contraceptives (years) | IUD | ... | STDs: Time since first diagnosis | STDs: Time since last diagnosis | Dx:Cancer | Dx:CIN |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 18 | 4.0 | 15.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | ? | ? | 0 | 0 |
| **1** | 15 | 1.0 | 14.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | ? | ? | 0 | 0 |
| **2** | 34 | 1.0 | ? | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | ? | ? | 0 | 0 |
| **3** | 52 | 5.0 | 16.0 | 4.0 | 1.0 | 37.0 | 37.0 | 1.0 | 3.0 | 0.0 | ... | ? | ? | 1 | 0 |
| **4** | 46 | 3.0 | 21.0 | 4.0 | 0.0 | 0.0 | 0.0 | 1.0 | 15.0 | 0.0 | ... | ? | ? | 0 | 0 |

5 rows × 36 columns


Age
Number of sexual partners
First sexual intercourse
Num of pregnancies
Smokes
Smokes (years)
Smokes (packs/year)
Hormonal Contraceptives
Hormonal Contraceptives (years)
IUD
IUD (years)
STDs
STDs (number)
STDs:condylomatosis
STDs:cervical condylomatosis
STDs:vaginal condylomatosis
STDs:vulvo-perineal condylomatosis
STDs:syphilis
STDs:pelvic inflammatory disease
STDs:genital herpes
STDs:molluscum contagiosum
STDs:AIDS
STDs:HIV
STDs:Hepatitis B
STDs:HPV
STDs: Number of diagnosis
STDs: Time since first diagnosis
STDs: Time since last diagnosis
Dx:Cancer
Dx:CIN
Dx:HPV
Dx
Hinselmann
Schiller
Citology
Biopsy
1
858
13
36


858 832 851 802 845 845 845 750 750 741 741 753 753 753 753 753 753 753 753 753 753 753 753 753 753 858 71 71 858 858 858 858 858 858 858 858
1.0
0.8
0.6
0.4
0.2
0.0
858
686
514
343
171
0
Age
Number of sexual partners
First sexual intercourse
Num of pregnancies
Smokes
Smokes (years)
Smokes (packs/year)
Hormonal Contraceptives
Hormonal Contraceptives (years)
IUD
IUD (years)
STDs
STDs (number)
STDs:condylomatosis
STDs:cervical condylomatosis
STDs:vaginal condylomatosis
STDs:vulvo-perineal condylomatosis
STDs:syphilis
STDs:pelvic inflammatory disease
STDs:genital herpes
STDs:molluscum contagiosum
STDs:AIDS
STDs:HIV
STDs:Hepatitis B
STDs:HPV
STDs: Number of diagnosis
STDs: Time since first diagnosis
STDs: Time since last diagnosis
Dx:Cancer
Dx:CIN
Dx:HPV
Dx
Hinselmann
Schiller
Citology
Biopsy



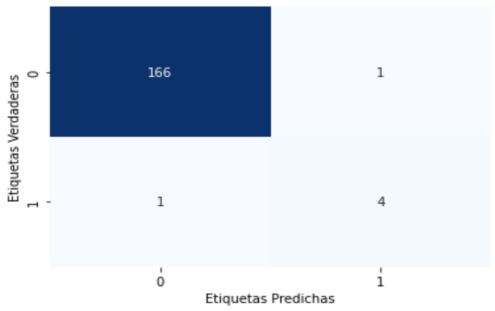

2. Imputación de datos perdidos
   Los valores perdidos se imputaron con la media de cada columna

3. Creación de un Modelo Base de Clasificación
   Se uso un RandomForest
   a) Se dividió los datos de características (X) y los datos de Objetivo (y='Dx:Cancer')
   b) Se dividieron los datos en conjuntos de entrenamiento y prueba. Proporción 80(training) -20 (test)
   c) Cálculo de exactitud entrenamiento y test
      El accuracy en train es: 1.0
      El accuracy en test es: 0.9883720930232558
      Se graficó los atributos según la relevancia en la predicción

| | 0 |
|---|---|
| **Dx:HPV** | 0.387782 |
| **Dx** | 0.199874 |
| **Age** | 0.061156 |
| **Hormonal Contraceptives (years)** | 0.043688 |
| **Num of pregnancies** | 0.037162 |
| **First sexual intercourse** | 0.030271 |
| **Smokes (packs/year)** | 0.029011 |
| **Smokes (years)** | 0.028547 |
| **STDs:HPV** | 0.024076 |
| **Citology** | 0.022755 |
| **IUD (years)** | 0.020288 |
| **Number of sexual partners** | 0.019236 |
| **Dx:CIN** | 0.017487 |
| **Biopsy** | 0.017011 |
| **Hormonal Contraceptives** | 0.012657 |
| **IUD** | 0.011113 |
| **Hinselmann** | 0.008946 |
| **Schiller** | 0.007017 |
| **STDs (number)** | 0.004389 |
| **STDs** | 0.004242 |

d) Informe de clasificación en cuanto a matriz de confusión y reporte de precisión accuracy y recall en test & training

```
Informe clasificacion en TEST
```

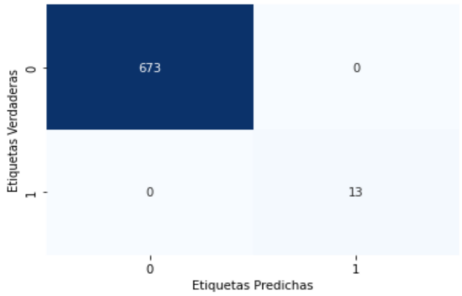



```
              precision    recall  f1-score   support

           0       0.99      0.99      0.99       167
           1       0.80      0.80      0.80         5

    accuracy                           0.99       172
   macro avg       0.90      0.90      0.90       172
weighted avg       0.99      0.99      0.99       172
```

```
Informe clasificacion en TRAINING
```




| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 1.00 | 1.00 | 1.00 | 673 |
| 1 | 1.00 | 1.00 | 1.00 | 13 |
| accuracy | | | 1.00 | 686 |
| macro avg | 1.00 | 1.00 | 1.00 | 686 |
| weighted avg | 1.00 | 1.00 | 1.00 | 686 |

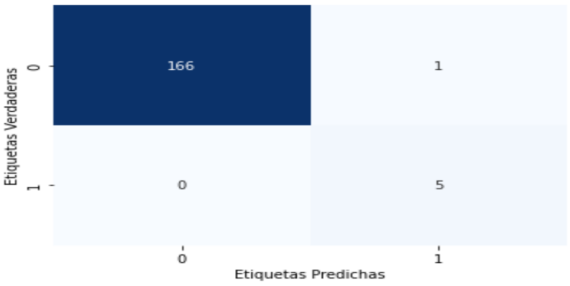
Los resultados son bien buenos en el modelo base

4. Dado el problema de desbalanceado en las clases objetivos,  se planteó hacer unos segundos modelos para remuestrear los datos de la clase objetivo y ver si hay mejoras.

**Opcion 1. OverSampling**

En este ejemplo, se utiliza la biblioteca imbalanced-learn (imblearn) que proporciona implementaciones de técnicas de remuestreo para abordar el desbalanceo de clases. En particular, se utiliza SMOTE para generar ejemplos sintéticos de la clase minoritaria y equilibrar las clases en el conjunto de entrenamiento

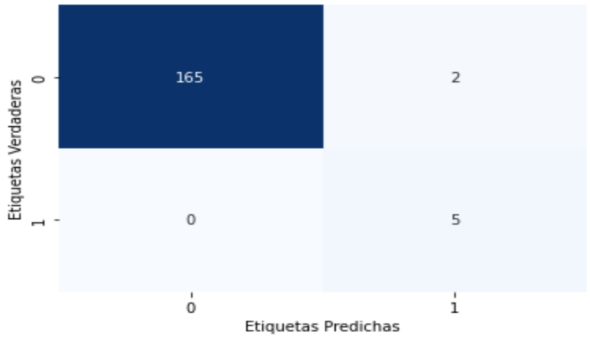
```
El accuracy en train es: 1.0
El accuracy en test es: 0.9941860465116279
```




| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 1.00 | 0.99 | 1.00 | 167 |
| 1 | 0.83 | 1.00 | 0.91 | 5 |
| accuracy | | | 0.99 | 172 |
| macro avg | 0.92 | 1.00 | 0.95 | 172 |
| weighted avg | 1.00 | 0.99 | 0.99 | 172 |

## Opción 2. UnderSampling- Reducir la clase mayoritaria

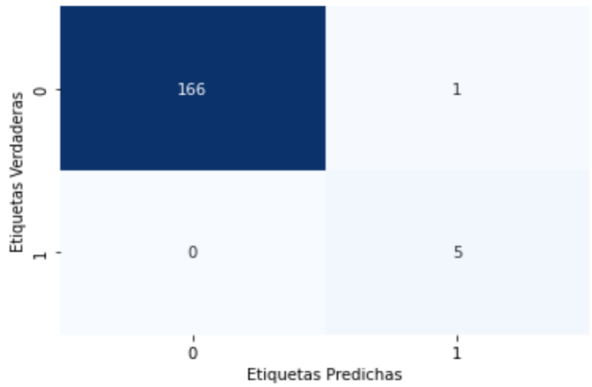
```
El accuracy en train es: 0.9927113702623906
El accuracy en test es: 0.9883720930232558
```




| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 1.00 | 0.99 | 0.99 | 167 |
| 1 | 0.71 | 1.00 | 0.83 | 5 |
| accuracy | | | 0.99 | 172 |
| macro avg | 0.86 | 0.99 | 0.91 | 172 |
| weighted avg | 0.99 | 0.99 | 0.99 | 172 |

## Opcion 3. un algoritmo de subsampling y otro de oversampling a la vez al dataset

```
El accuracy en train es: 1.0
El accuracy en test es: 0.9941860465116279
```




| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 1.00 | 0.99 | 1.00 | 167 |
| 1 | 0.83 | 1.00 | 0.91 | 5 |
| accuracy | | | 0.99 | 172 |
| macro avg | 0.92 | 1.00 | 0.95 | 172 |
| weighted avg | 1.00 | 0.99 | 0.99 | 172 |

Como conclusion. El modelo tiene mejor clasificación en ambas clases con el modelo de oversampling y subsampling con oversatting al tiempo, en especial para la clase l 1.