MU BIOINFORMÁTICA Y BIOESTADÍSTICA

RAMA BIOINFORMÁTICA ESTADÍSTICA Y APRENDIZAJE AUTOMÁTICO

Desarrollo de una aplicación web para la predicción de partos prematuros, aplicando técnicas de aprendizaje automático sobre las características etiológicas de mujeres embarazadas de India

Claudia Francesca Llinares Monllor

Tutora de TF: Romina Astrid Rebrij

Universitat Oberta de Catalunya

Universitat de Barcelona

Índice

Introducción: Datos generales

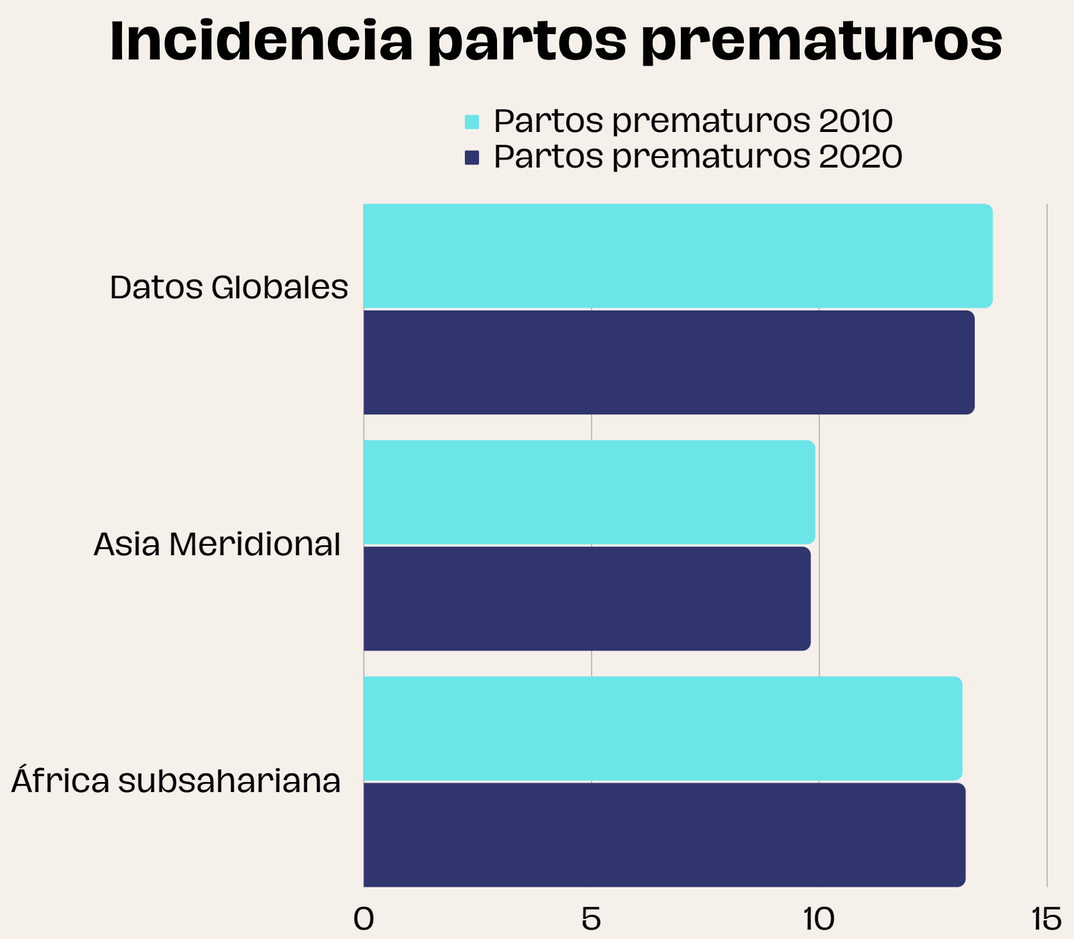
Partos prematuros: ¿problema de salud mundial?

- Los partos prematuros y las complicaciones derivadas de ello se han postulado como uno de los grandes problemas a nivel de salud mundial.

- Del 2010 al 2020 se calcula que han habido 152 millones de bebés vulnerables debido a un nacimiento prematuro.

- En esta década, no ha habido cambios de tendencias:
  - Nivel global: De 9,9% a 9,8% de partos prematuros
  - Asia Meridional: 13.13%a 13.12% de partos prematuros
  - África subsahariana: 10.1% en la década

- Existen excepciones de países que sí reducen estos niveles (España), pero otros lo incrementan (Islandia). No obstante, se pasó de 13.8 a 13.4 millones de partos prematuros.



Fuente bibliográfica: (1)

Introducción: Datos generales

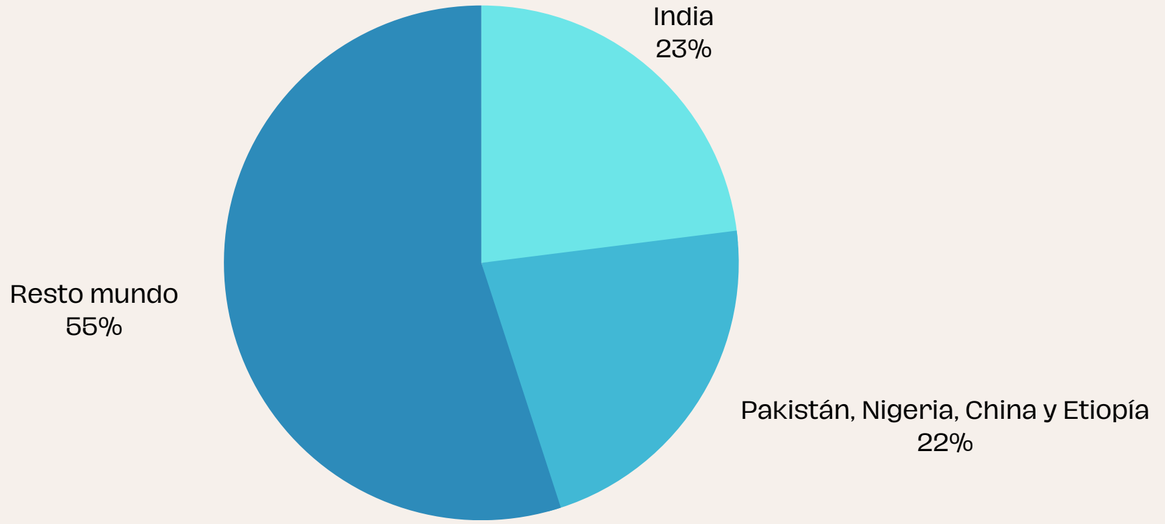
Partos prematuros: ¿problema de salud mundial?

- La distribución de partos prematuros a nivel mundial varía dependiendo de la zona y el país.

- Mayor incidencia en sur de Asia, el 45% de los partos prematuros ocurren en:
  - India
  - Pakistán
  - Nigeria
  - China
  - Etiopía

- India es el país con más partos prematuros en 2020, con 3.02 millones de nacimientos prematuros.

Distribución mundial parto prematuro



Fuente bibliográfica (1)

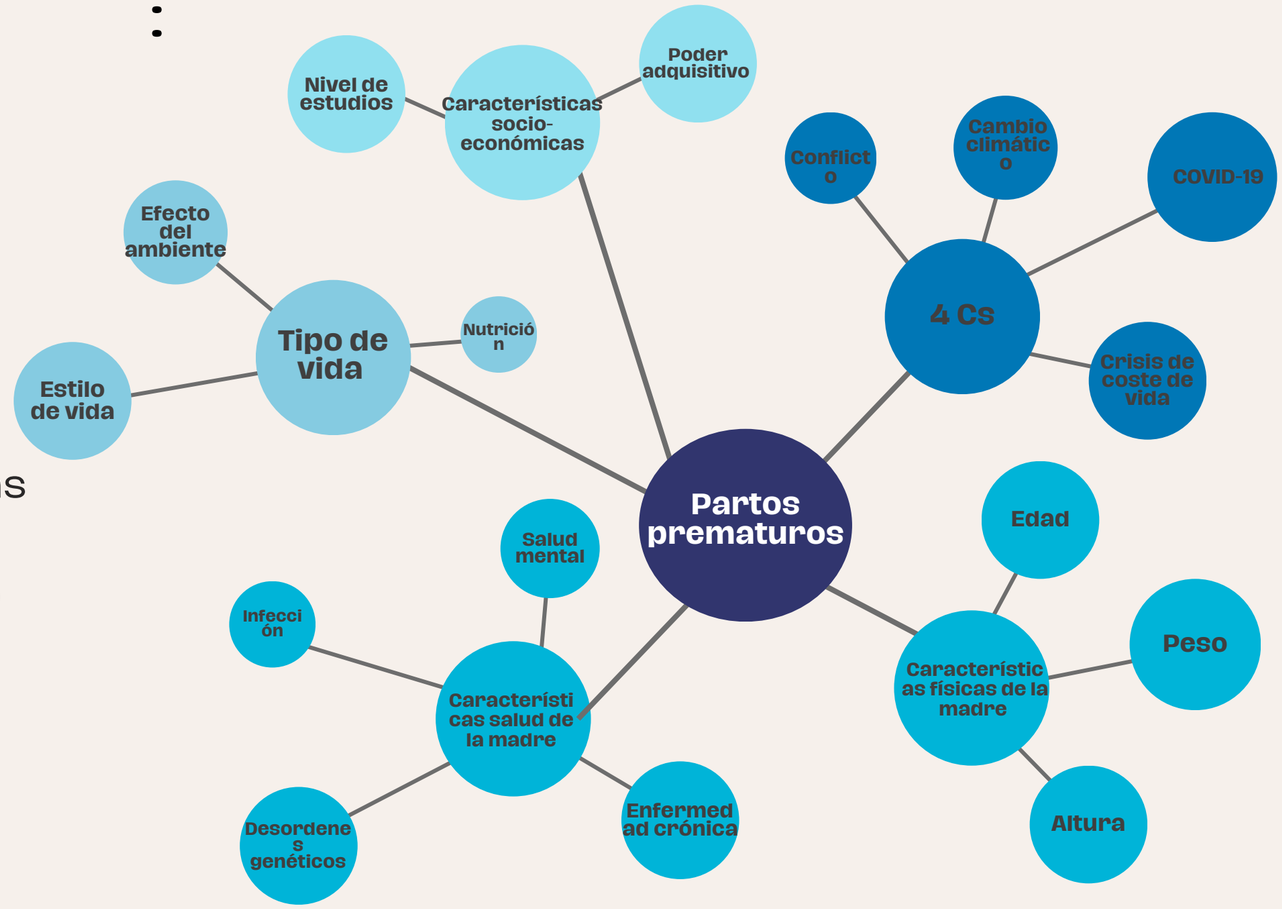
Introducción: Causas

Partos prematuros: ¿problema de salud mundial?

- Los nuevos informes de la OMS y NU ponen en el foco cuatro problemas:
  - Conflicto
  - Cambio climático
  - COVID-19
  - Crisis de coste de vida

- Nuevos estudios también determinan nuevas causas relacionadas con diferentes aspectos de la madre:
  - Características físicas de la madre: edad o peso
  - Características salud de la madre: Infecciones, salud mental o desordenes genéticos
  - Características tipo de vida de la madre: Estilo de vida o efecto del ambiente
  - Características socio-económicas: Poder adquisitivo o nivel de estudios

Causas:



Fuente bibliográfica: (2)

Introducción: Efectos

Partos prematuros: ¿problema de salud mundial?

Infantes

- Los partos prematuros constituyen la cuarta causa principal de pérdida de capital humano en todo el mundo.
  - Son la primera causa de mortalidad infantil
- Casi un millón de niños mueren por las complicaciones en nacimientos prematuros.
- Un tercio de los 2.3 millones de muertes prematuras son como consecuencia de los nacimientos prematuros.

Madres

- De 4.5 millones de muertes al año debido a partos prematuros, 287.000 corresponden al fallecimiento de madres.
- A pesar de programas específicos de varias organizaciones, el decrecimiento de problemas en el parto se estancó en 133 países.
- Existen desigualdades en tratamiento y seguimiento durante el embarazo tanto a nivel de países como dentro del mismo país.

Fuente bibliográfica: (3,4)

Introducción: Soluciones

Partos prematuros: ¿Existen soluciones?

Soluciones actuales:

- Acelerar o mejorar desarrollo de los bebés
- Retrasar el parto prematuro

Problemas:

- Existen desigualdades socio-económicas
- Recursos inaccesibles en ciertas zonas del mundo

→

¿Nuevas soluciones?:

- Prevención primaria (seguimiento durante el embarazo)
- Estudio de la madre

Abordaje:

- Descripción del perfil completo de la madre
- Digitalización del sistema de salud

→

Hipótesis de trabajo:

- Determinación de factores influyentes en el parto prematuro
- Estudio de predicción

Metodología:

- Conjunto de datos MSF
- Análisis de datos
- Desarrollo aplicación web

Fuente bibliográfica: (5,6,7)

Objetivos

Generales

- Estudio de las múltiples variables dentro del conjunto de datos escogido para conocer cuáles tienen mayor efecto en el parto prematuro y determinación de un perfil multifactorial específico para la predicción de partos prematuros.

- Desarrollo de un modelo de aprendizaje automático capaz de predecir si un embarazo tiene riesgos en acabar de forma prematura y creación de una aplicación web que permita predecir si un parto es prematuro o no con la inclusión de nuevos datos.

Específicos

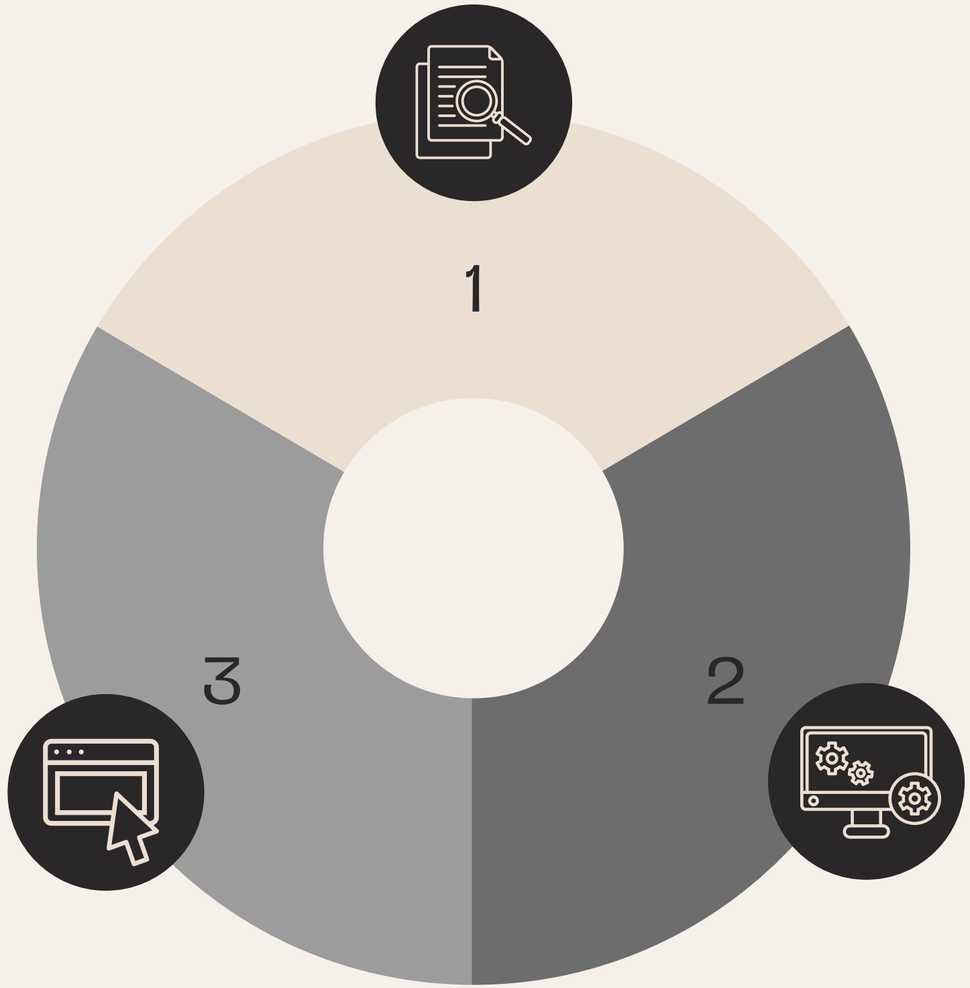
- Realizar un análisis exploratorio de las 150 variables del conjunto de datos original
- Usar herramientas de reducción de dimensionalidad para escoger las mejores variables
- Crear un conjunto de datos final con las variables más óptimas

- Evaluar diversas técnicas de clasificación para determinar el mejor modelo predictivo
- Mejorar y optimizar los modelos hasta obtener una precisión mínima del 75-80%
- Desarrollar una aplicación web en base al mejor modelo obtenido con nuestros datos

Materiales y métodos

1.ANÁLISIS DE DATOS

- Análisis exploratorio de datos
- Estudio de valores nulos
- Estudio de valores atípicos
- Reducción dimensionalidad



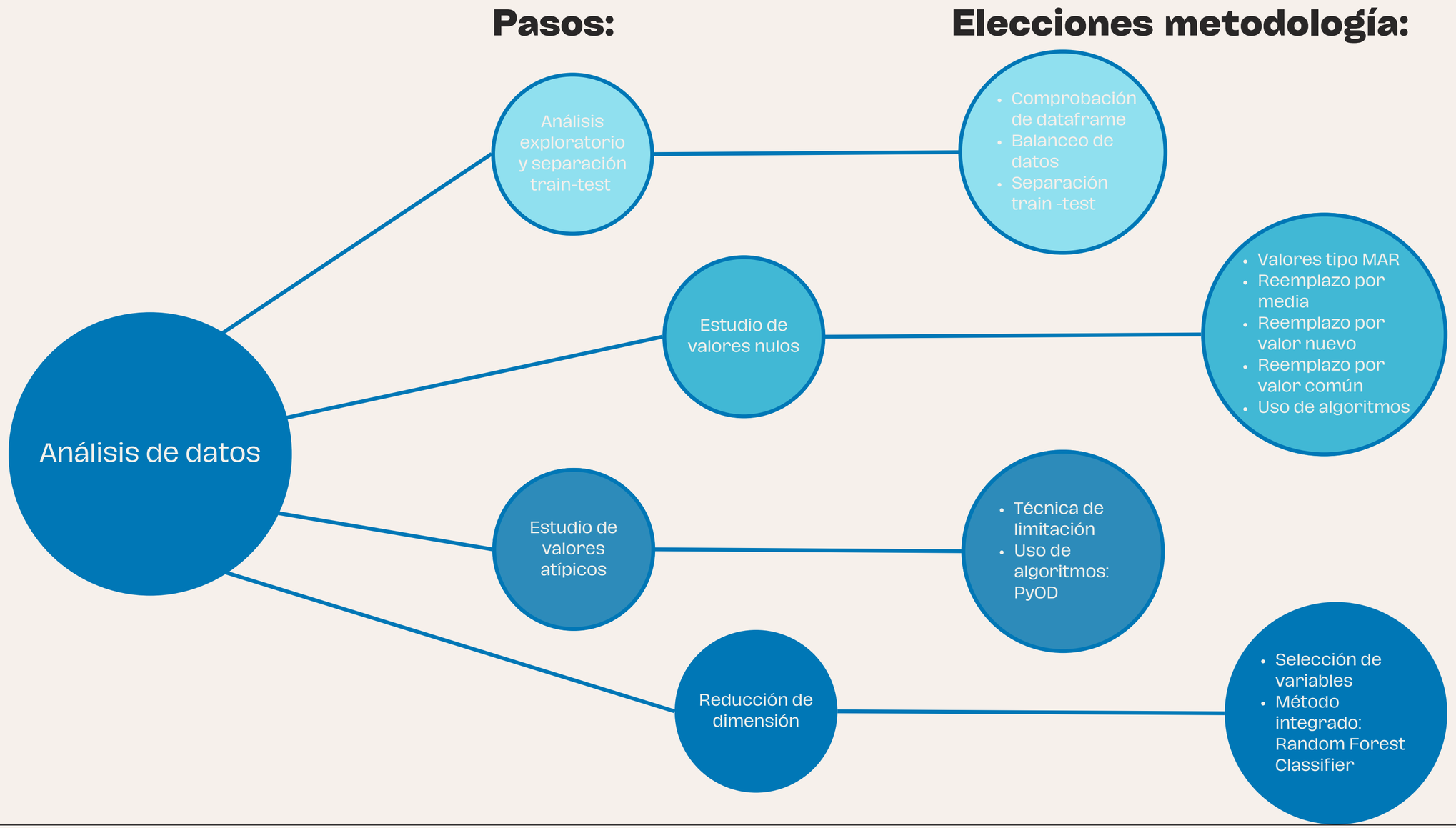
2. APRENDIZAJE AUTOMÁTICO

- Árboles de decisión
- Bosques aleatorios
- Máquinas de vectores de soporte
- Redes neuronales artificiales

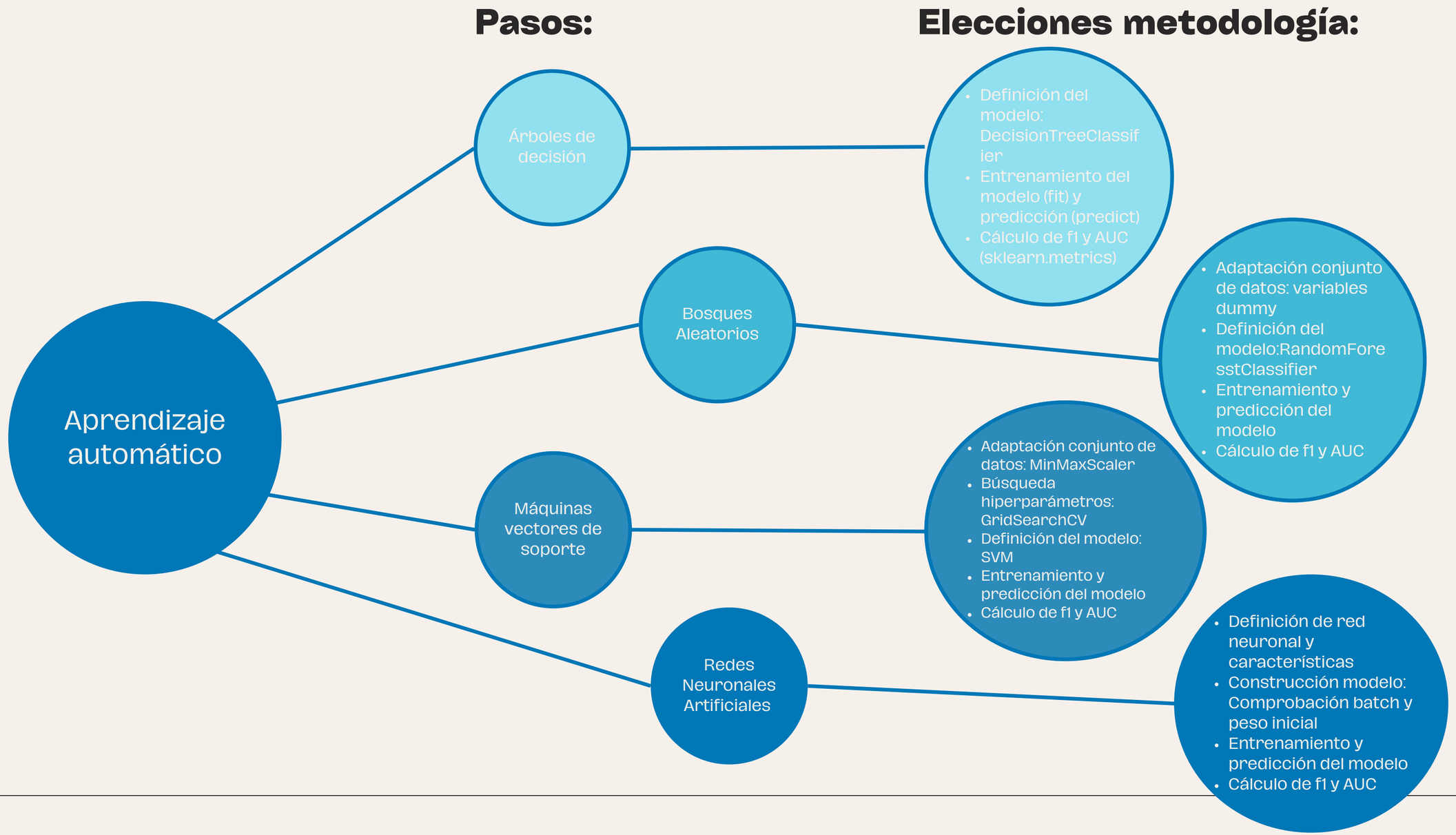
3.PRODUCTOS

- Aplicación web
- Repositorio público

Materiales y métodos (1)



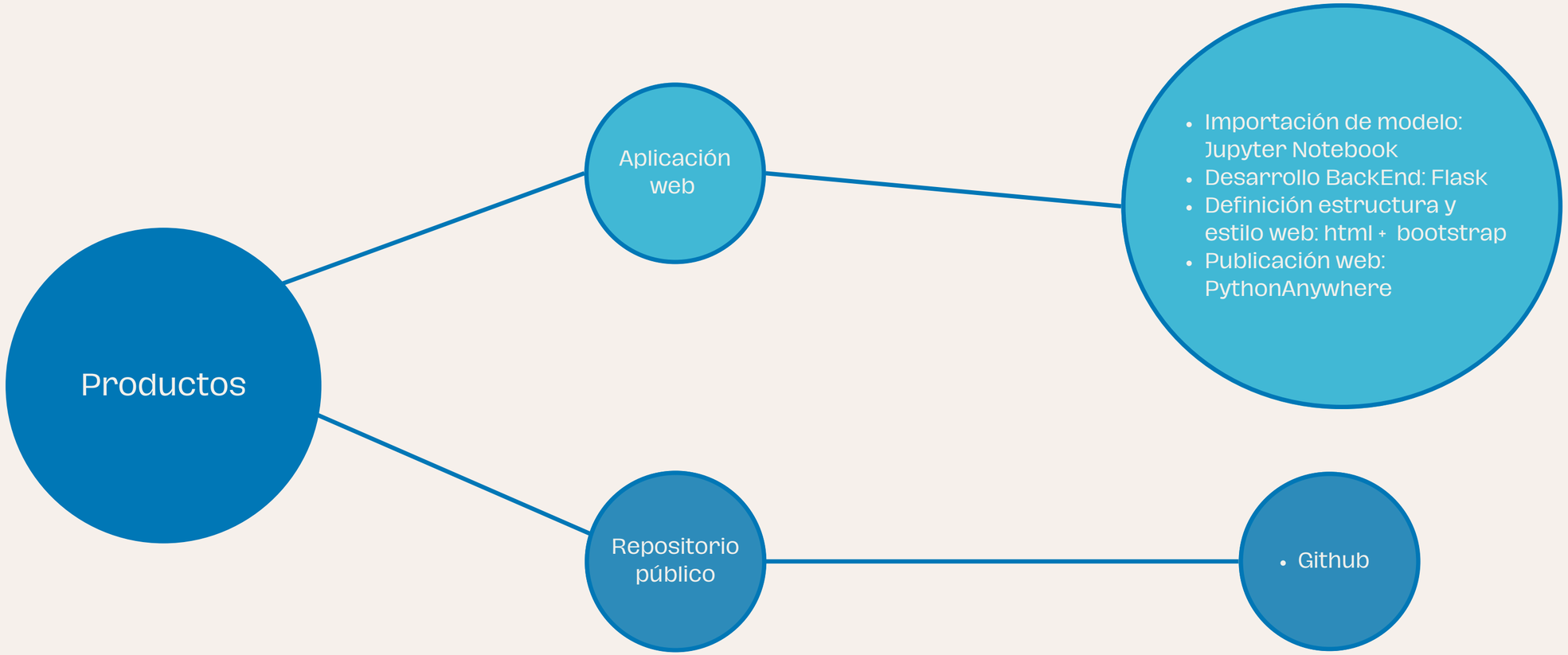
Materiales y métodos (2)



Materiales y métodos (3)

Pasos:

Elecciones metodología:



Resultados

1. Análisis de datos

Análisis exploratorio y separación entrenamiento - test

Análisis exploratorio:

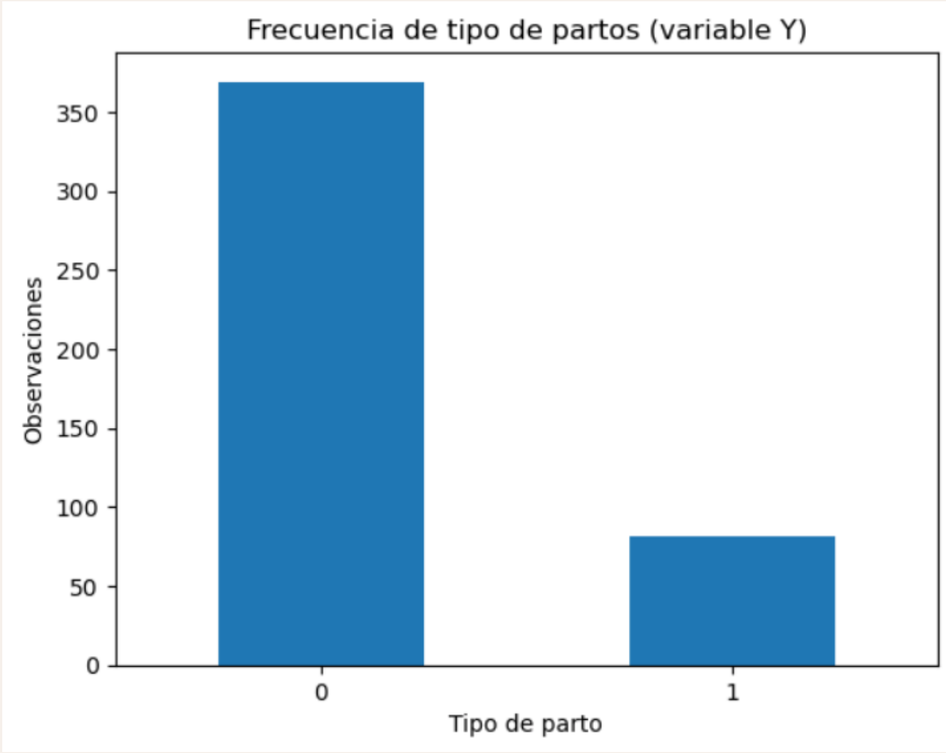
```
# Con el comando .info() podemos obtener un resumen general de las
# características de esta base de datos.
datos.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 450 entries, 0 to 449
Columns: 131 entries, Mother_UID to Induce_Pain
dtypes: float64(24), int64(107)
memory usage: 460.7 KB
```

Separación train-test:

```
# Dividimos el data set en conjunto de entrenamiento y conjunto de testing usando
# la función de train_test_split del paquete de sklearn:
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size = 0.2,
                                                    random_state = 123, stratify = Y['PreTerm'])
```

Estudio frecuencia tipo partos:



Resultados

1. Análisis de datos

Preprocesamiento: Estudio de valores nulos

Identificación variables con NA:

```
# Mostramos el nombre de las columnas con valores NA para X_test
X_test.columns[X_test.isna().any()]
```

```
Index(['wt_before_delivery', 'No_of_sibling', 'Miscarrage History',
       'Food_crav', 'Mood_swing', 'Wishing_vac', 'Visiting places',
       'Artistic things', 'Shopping', 'Cooking', 'Spending_time_people',
       'Eating', 'Sitting_alone', 'LeaveWork_mon', 'Family_Support_inlaws',
       'Family_support_parents', 'Family_support_husband',
       'You_Support_inlaws', 'You_support_parents', 'You_support_other',
       'Diabetes_preg', 'None'],
      dtype='object')
```

Tratamiento variable numérica continua:

```
# Tratamiento NA para variable wt_before_delivery en X_train:
meanwt = X_train['wt_before_delivery'].mean()
X_train['wt_before_delivery'].fillna(meanwt, inplace=True)

# Comprobamos si ha ido bien y hemos eliminado los valores NA:
X_train['wt_before_delivery'].isna().sum()
```

```
0
```

Tratamiento variable numérica discontinua:

```
# Tratamiento NA para para la variable Miscarrage History en X_train:
X_train['Miscarrage History']= X_train['Miscarrage History'].fillna(X_train[
    'Miscarrage History'].mode().iloc[0])

# Comprobamos si ha ido bien y hemos eliminado los valores NA:
X_train['Miscarrage History'].isna().sum()
```

```
0
```

Resultados

1. Análisis de datos

Preprocesamiento: Estudio de valores nulos

Tratamiento variable categórica 1:

```
# Sustitución de NAs por nueva clase - X_train:
X_train1 = X_train.fillna(10)

# Comprobamos que se haya hecho la sustitución de manera correcta
X_train1.head()
```

	Mother_UID	Age_Of_Mother	weight_before_preg	wt_before_delivery	Height(cm)	BMI	Hem
247	248	27	55	64.426752	142	28	
249	250	31	48	64.426752	142	24	
185	186	29	72	74.000000	172	25	
323	324	21	60	64.426752	149	28	
80	81	32	78	82.000000	179	25	

Tratamiento variable categórica 2:

```
# Sustitución de NAs por valor más común - X_train:
X_train2 = X_train.apply(lambda x: x.fillna(x.value_counts().index[0]))

# Comprobamos que se haya hecho la sustitución de manera correcta
X_train2.head()
```

	Mother_UID	Age_Of_Mother	weight_before_preg	wt_before_delivery	Height(cm)	BMI	Hem
247	248	27	55	64.426752	142	28	
249	250	31	48	64.426752	142	24	
185	186	29	72	74.000000	172	25	
323	324	21	60	64.426752	149	28	
80	81	32	78	82.000000	179	25	

Tratamiento variable categórica 3:

```
# 3º método: KNN neighborn - X_train:
# Cargamos la librería y paquetes específicos para la imputación media
from sklearn.impute import KNNImputer

# Definimos las características del algoritmo (3 vecinos):
knni = KNNImputer(n_neighbors=3)

# Aplicamos el algoritmo a nuestro conjunto de datos X_train:
X_train_knn = knni.fit_transform(X_train)

# Transformamos el resultado en dataframe:
X_train3 = pd.DataFrame(np.round(X_train_knn), columns = X_train.colum

# Comprobamos que se hayan eliminado las valores NA:
X_train3.isnull().any().sum()
```

```
0
```

Resultados

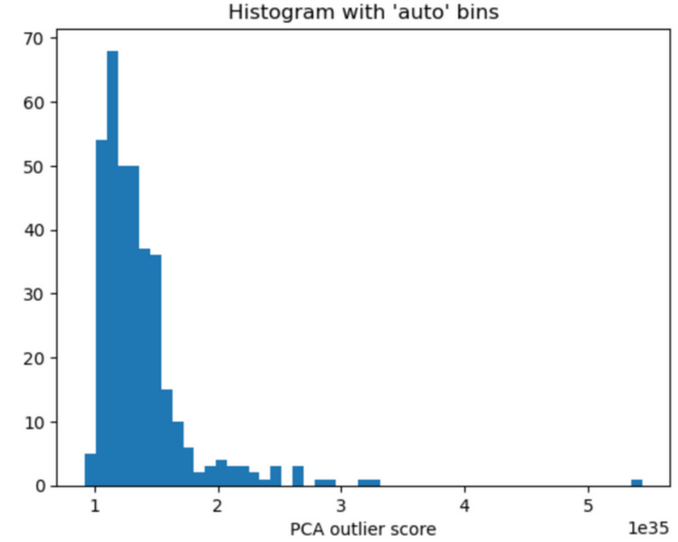
1. Análisis de datos

Preprocesamiento: Estudio de valores atípicos _ PCA

PASOS:

- Desarrollo del modelo
- Cálculo de treshold
- Determinación de grupos normal y outlier

Conjunto de datos 1




```
# Definimos la threshold de este modelo:
print("El treshold para el ratio de contaminación de este modelo es:", pca1.threshold_)
```

```
El treshold para el ratio de contaminación de este modelo es: 1.6903918644462668e+35
```

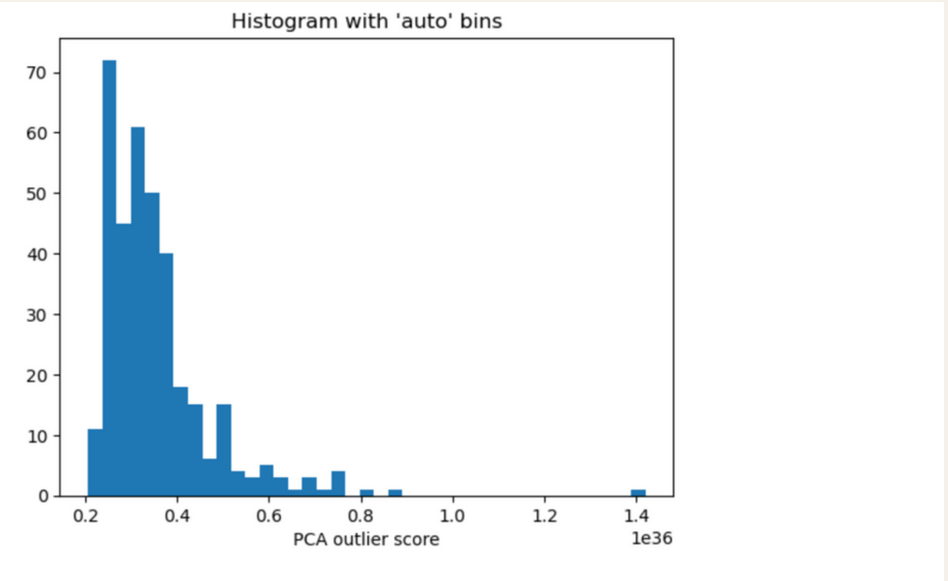
	Group	Count	Count %	Mother_UID	Age_Of_Mother	weight_before_preg	wt_before_delivery
0	Normal	324	90.0	232.29	26.30	55.88	64.00
1	Outlier	36	10.0	177.53	29.44	63.42	68.29

Resultados

1. Análisis de datos

Preprocesamiento: Estudio de valores atípicos _ PCA

Conjunto de datos 2

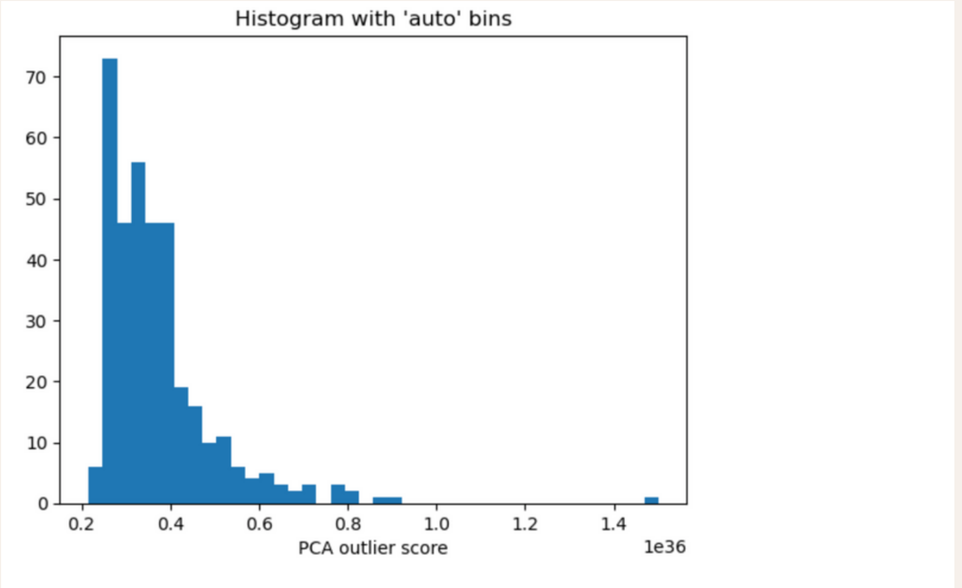



```
# Definimos la threshold de este modelo:
print("El treshold para el ratio de contaminación de este modelo es:", pca2.threshold_)
```

```
El treshold para el ratio de contaminación de este modelo es: 4.988273645274809e+35
```

	Group	Count	Count %	Mother_UID	Age_Of_Mother	weight_before_preg	wt_before_delivery
0	Normal	324	90.0	235.54	26.46	56.11	64.21
1	Outlier	36	10.0	148.28	28.06	61.33	66.41

Conjunto de datos 3




```
# Definimos la threshold de este modelo:
print("El treshold para el ratio de contaminación de este modelo es:", pca3.threshold_)
```

```
El treshold para el ratio de contaminación de este modelo es: 5.2903430058472606e+35
```

	Group	Count	Count %	Mother_UID	Age_Of_Mother	weight_before_preg	wt_before_delivery
0	Normal	324	90.0	235.52	26.43	56.07	63.92
1	Outlier	36	10.0	148.53	28.31	61.69	66.61

Resultados

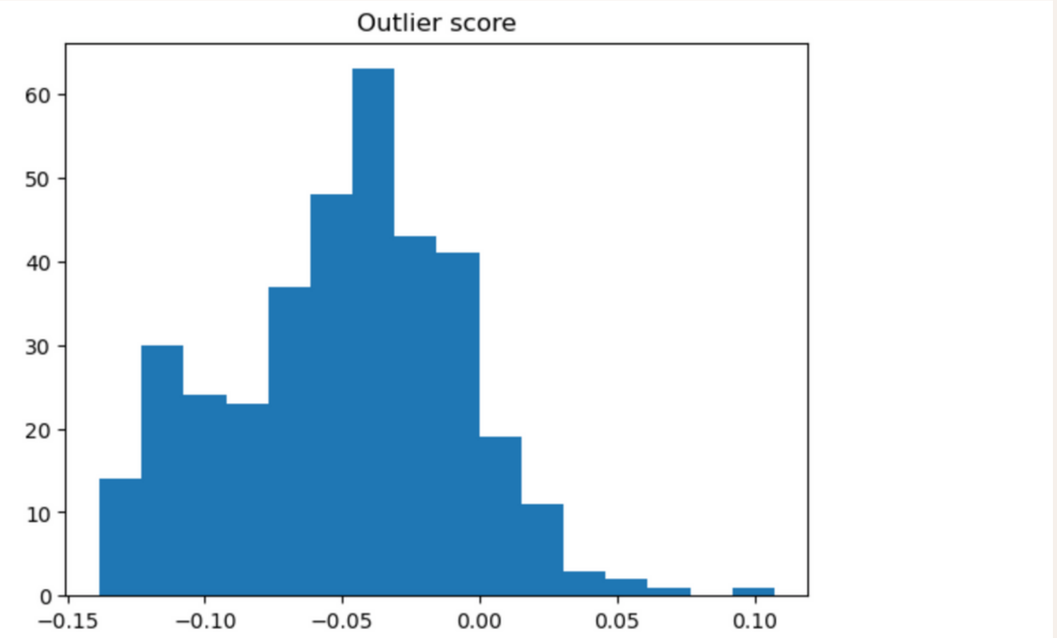
1. Análisis de datos

Preprocesamiento: Estudio de valores atípicos _ Isolation Forest

PASOS:

- Desarrollo del modelo
- Cálculo de treshold
- Determinación de grupos normal y outlier

Conjunto de datos 1




```
# Definimos la threshold de este modelo:
print("The threshold for the defined contamination rate:" , isftI1.threshold_)
```

```
The threshold for the defined contamination rate: 4.449527599397884e-17
```

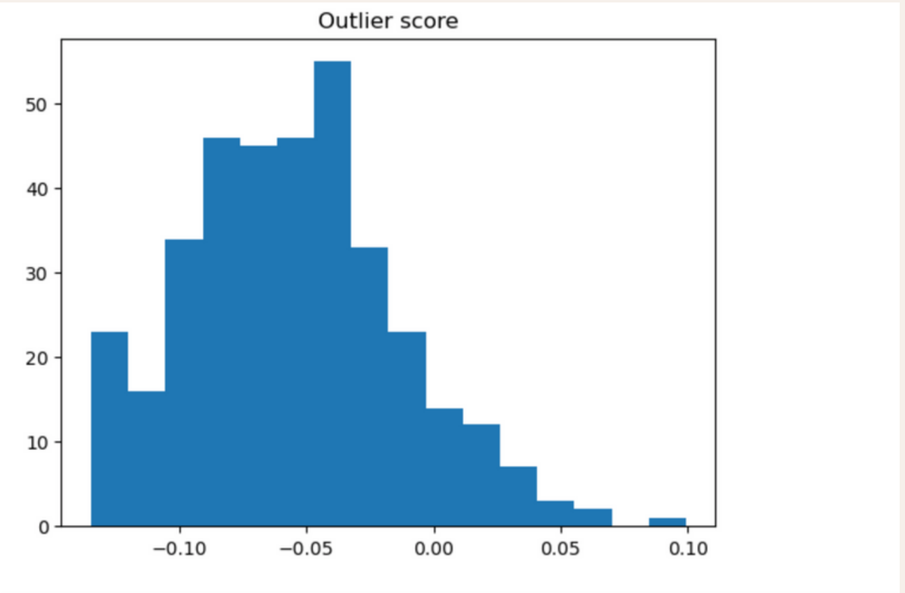
	Group	Count	Count %	Mother_UID	Age_Of_Mother	weight_before_preg	wt_before_delivery
0	Normal	324	90.0	230.52	26.26	56.10	64.14
1	Outlier	36	10.0	193.47	29.83	61.44	66.98

Resultados

1. Análisis de datos

Preprocesamiento: Estudio de valores atípicos _ Isolation Forest

Conjunto de datos 2

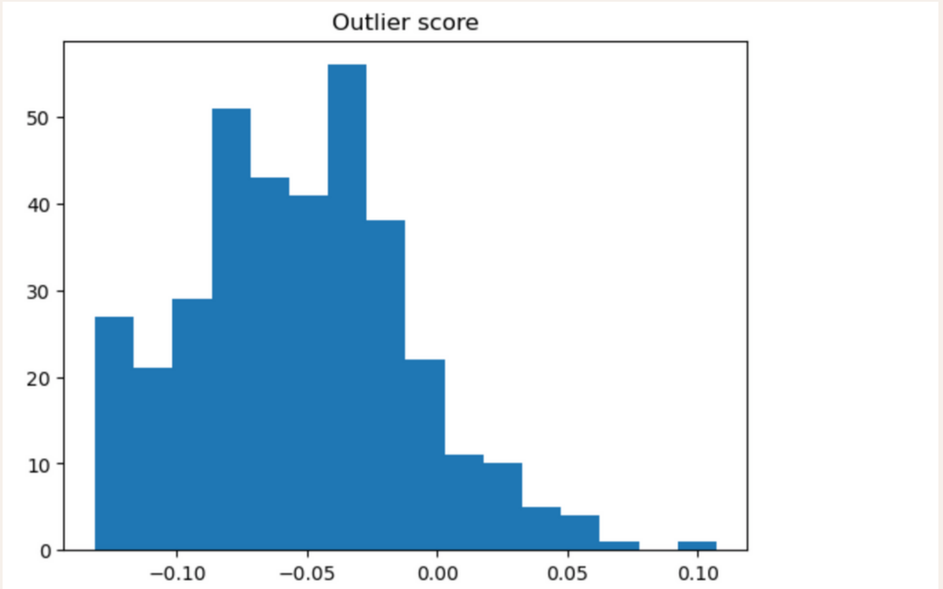



```
# Definimos la threshold de este modelo:
print("The threshold for the defined contamination rate:" , isftI2.threshold_)
```

```
The threshold for the defined contamination rate: 2.6490955425378243e-19
```

	Group	Count	Count %	Mother_UID	Age_Of_Mother	weight_before_preg	wt_before_delivery
0	Normal	324	90.0	240.43	26.30	55.77	64.08
1	Outlier	36	10.0	104.31	29.44	64.36	67.55

Conjunto de datos 3




```
# Definimos la threshold de este modelo:
print("The threshold for the defined contamination rate:" , isftI3.threshold_)
```

```
The threshold for the defined contamination rate: -7.592803339187548e-18
```

	Group	Count	Count %	Mother_UID	Age_Of_Mother	weight_before_preg	wt_before_delivery
0	Normal	324	90.0	238.92	26.34	55.94	63.97
1	Outlier	36	10.0	117.92	29.08	62.81	66.14

Resultados

1. Análisis de datos

Preprocesamiento: Estudio de valores atípicos _ Comparación modelos

Matriz de confusión

Detección valores atípicos: matriz de confusión

(A)

Actual \ Pred	0	1
0	264	31
1	60	5

(B)

Actual \ Pred	0	1
0	265	30
1	59	6

(C)

Actual \ Pred	0	1
0	265	30
1	59	6

(D)

Actual \ Pred	0	1
0	267	28
1	57	8

(E)

Actual \ Pred	0	1
0	265	30
1	59	6

(F)

Actual \ Pred	0	1
0	265	30
1	59	6

Valores de precisión

```
Accuracy pca1 74.72222222222223%
Accuracy pca2 75.27777777777777%
Accuracy pca3 75.27777777777777%
Accuracy IF1 76.38888888888889%
Accuracy IF2 75.27777777777777%
Accuracy IF3 75.27777777777777%
```

INTERPRETACIÓN:

- Valores de la matriz de confusión con pocos TP y TN.
- Valores precisión superiores a 74%
- No se eliminan los valores atípicos

Resultados

1. Análisis de datos

Preprocesamiento: Reducción de dimensionalidad

Reducción dimensionalidad: tabla de variables

(A)

	Importance	Features
5	0.051628	Hemoglobin
0	0.051433	Age_Of_Mother
7	0.045655	Age_Father
1	0.045599	weight_before_preg
3	0.042384	Height(cm)
4	0.038190	BMI
8	0.037791	Yrs_Of_Marriage
113	0.026770	Gastric_preg
12	0.026146	Education
119	0.025428	no of births(single/Twins)

(B)

	Importance	Features
0	0.051478	Age_Of_Mother
5	0.049506	Hemoglobin
7	0.043230	Age_Father
1	0.042923	weight_before_preg
3	0.041886	Height(cm)
8	0.040464	Yrs_Of_Marriage
4	0.034256	BMI
12	0.026237	Education
113	0.024982	Gastric_preg
109	0.019720	Family_Income

(C)

	Importance	Features
7	0.050273	Age_Father
0	0.050102	Age_Of_Mother
3	0.045472	Height(cm)
8	0.040412	Yrs_Of_Marriage
4	0.036833	BMI
1	0.035650	weight_before_preg
5	0.028796	Hemoglobin
12	0.026583	Education
119	0.026567	no of births(single/Twins)
113	0.024132	Gastric_preg

INTERPRETACIÓN:

- Se comparten 9 de las 10 variables en los tres conjuntos de datos
- Se crean 3 nuevos conjuntos de datos

Resultados

2. Aprendizaje automático

Árboles de decisión

	X_train1d	X_train2d	X_train3d
Valor f1	0.7972	0.7262	0.6986
Valor AUC	0.5979	0.5194	0.5346

Bosques aleatorios

	ohe_data1	ohe_data2	ohe_data3
Valor f1	0.7421	0.7420	0.7420
Valor AUC	0.5	0.5	0.5

Resultados

2. Aprendizaje automático

Máquinas de vectores de soporte

	df_scaled1	df_scaled2	df_scaled3
Valor f1	0.7718	0.6269	0.7636
Valor AUC	0.5954	0.5473	0.5887

Redes neuronales artificiales

	df_scaled1	df_scaled2	df_scaled3
Valor f1	0.3209	0.3209	0.3209
Valor AUC	0.5409	0.5409	0.5409

Resultados

2. Aprendizaje automático _ Optimización modelos

Comparaciones modelos:

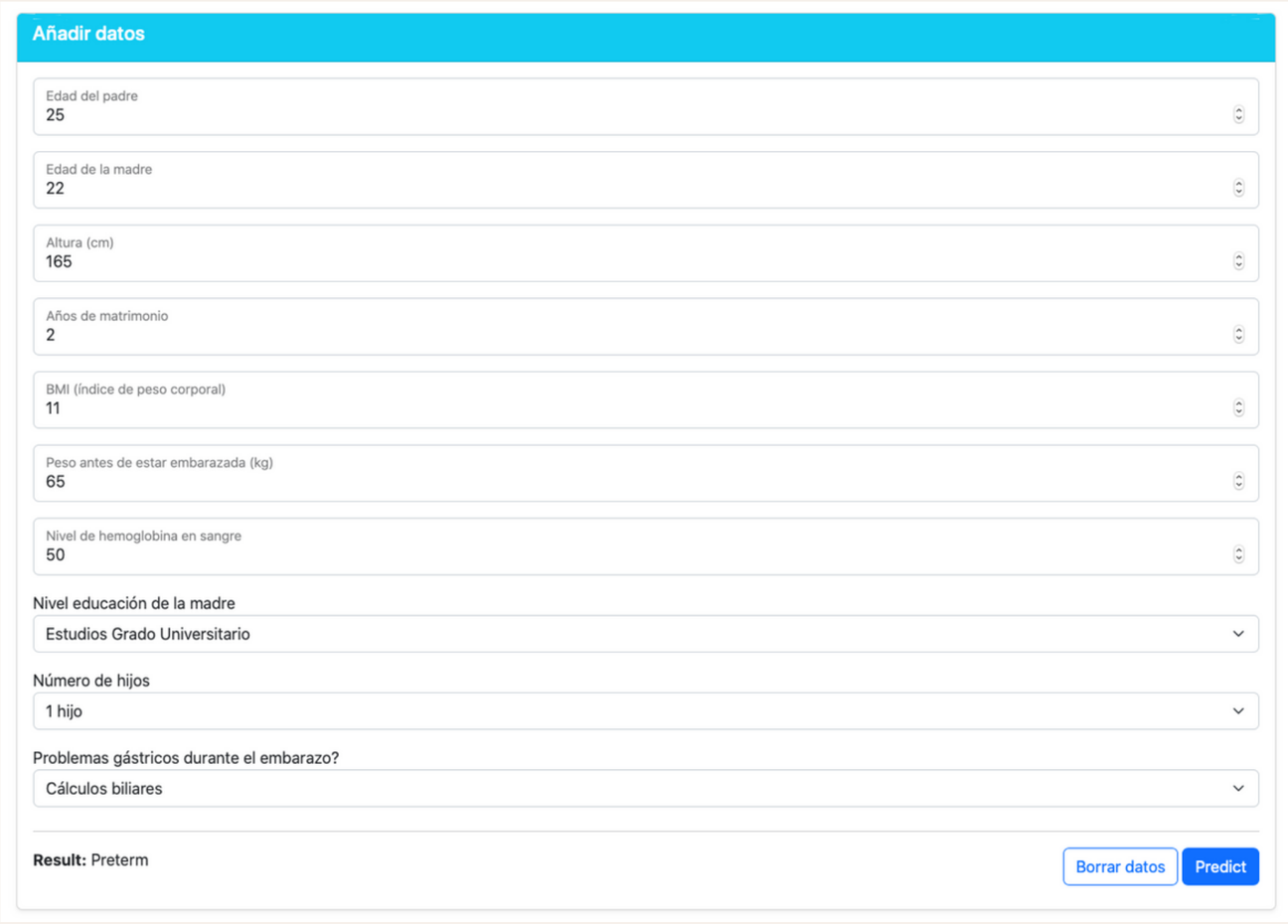
	Árboles de decisión (optimizado)	Bosques aleatorio (optimizado)	Máquinas de vectores de soporte	Redes neuronales artificiales
Conjunto de datos 1	0.8140	0.8036	0.7718	0.3209
Conjunto de datos 2	0.7755	0.7681	0.6269	0.3209
Conjunto de datos 3	0.8420	0.7681	0.7636	0.3209

INTERPRETACIÓN:

- Se comparan valores f1 de los tres conjuntos de datos
- El modelo con más valor es Árboles de decisión en el conjunto de datos 3

Resultados

3. Aplicación web



Link acceso:
https://claudiauoc.pythonanywhere.com

Conclusiones

Conclusiones:

Primer objetivo:

- Perfil multifactorial obtenido: Coinciden 9 de las 10 variables en los tres conjuntos de datos
- 6 de las 10 variables son características físicas de la madre
- Nivel de educación o ingresos mensuales

Segundo objetivo:

- Mejor método sustitución de valores nulos: imputación por algoritmo kNN
- Mejor modelo: árboles de decisión

Seguimiento planificación:

- Cumplimiento de todos los objetivos
- Modificaciones y/o ampliación de pasos = desajuste temporal
- Cambio de metodología: reducción de dimensionalidad

Líneas de futuro:

- Ampliación de variables en el estudio o cuestionario
- Estudio sobre impacto de factores climáticos en partos prematuros
- Exploración de más modelos de predicción centrados en una tipología de datos

Referencias bibliográficas

(1) Born too soon: decade of action on preterm birth. Geneva: World Health Organization; 2023. Licence: CC BY-NC-SA 3.0 IGO. (https://creativecommons.org/licenses/ by-nc-sa/3.0/igo/)

(2) World Health Organization, United Nations Children's Fund (UNICEF). Protect the Promise: 2022 progress report on the Every Woman Every Child Global Strategy for Women's, Children's and Adolescents' Health (2016–2030). Geneva: World Health Organization; 2022. (https://apps.who.int/iris/handle/10665/363919)

(3) Perin J, Mulick A, Yeung D, Villavicencio F, Lopez G, Strong KL, et al. Global, regional, and national causes of under-5 mortality in 2000-19: an updated systematic analysis with implications for the Sustainable Development Goals. Lancet Child Adolesc Health. 2022;6(2):106-15.

(4) Trends in maternal mortality 2000 to 2020: estimates by WHO, UNICEF, UNFPA, World Bank Group and UNDESA/ Population Division. Geneva: World Health Organization; 2023. (https://apps.who.int/iris/handle/10665/366225)

(5) Liu G, Segrè J, Gülmezoglu AM, Mathai M, Smith JM, Hermida J, et al. Antenatal corticosteroids for management of preterm birth: a multi-country analysis of health system bottlenecks and potential solutions. BMC pregnancy and childbirth. 2015;15(2):1-16.

(6) Meis, P. J., Goldenberg, R. L., Mercer, B. M., Iams, J. D., Moawad, A. H., Miodovnik, M., Menard, M. K., Caritis, S. N., Thurnau, G. R., Bottoms, S. F., Das, A., Roberts, J. M., & McNellis, D. (1998). The preterm prediction study: risk factors for indicated preterm births. Maternal-Fetal Medicine Units Network of the National Institute of Child Health and Human Development. American Journal of Obstetrics and Gynecology, 178(3), 562–567. https://doi.org/10.1016/s0002-9378(98)70439-9

(7) Cobo, T., Kacerovsky, M., & Jacobsson, B. (2020). Risk factors for spontaneous preterm delivery. International Journal of Gynaecology and Obstetrics: The Official Organ of the International Federation of Gynaecology and Obstetrics, 150(1), 17–23. https://doi.org/10.1002/ijgo.13184

Muchas gracias por la atención

JUNIO → 2023