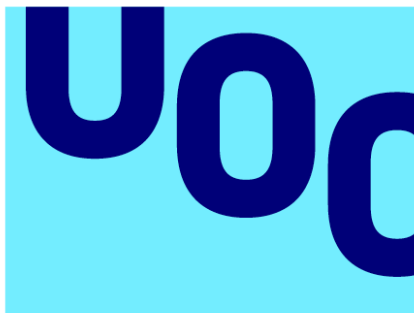


Desarrollo de una aplicación web para la predicción de partos prematuros, aplicando técnicas de aprendizaje automático sobre las características etiológicas de mujeres embarazadas de India



Universitat
Oberta
de Catalunya



UNIVERSITAT_{DE}
BARCELONA

Claudia Francesca Llinares Monllor

MU Bioinf. i Bioest.
Bioinformàtica Estadística y
Aprendizaje Automático

Tutor/a de TF

Romina Astrid Rebrij

**Professor/a responsable de
l'assignatura**

Carles Ventura Royo

20 de Junio 2023



Aquesta obra està subjecta a una llicència de [Reconeixement-NoComercial-SenseObraDerivada 3.0 Espanya](https://creativecommons.org/licenses/by-nc-nd/3.0/es/) de Creative Commons

FICHA DEL TRABAJO FINAL

Título del trabajo:	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
Nombre del autor:	Claudia Francesca Llinares Monllor
Nombre del consultor/a:	Romina Astrid Rebrij
Nombre del PRA:	Carles Ventura Royo
Fecha de entrega (mm/aaaa):	06/2023
Titulación o programa:	Máster Bioinformática y Bioestadística
Área del Trabajo Final:	Bioinformática Estadística y Aprendizaje Automático
Idioma del trabajo:	Castellano
Palabras clave	Parto prematuro, aprendizaje supervisado, aplicación web
Resumen del Trabajo	
<p>Los partos prematuros son un problema de gran importancia que afectan a todos los países del mundo. Es así que de 15 millones de nacimientos producidos al año, el 84% se consideran prematuros. No sólo implican un problema a nivel de complicaciones en el parto y el peligro que conlleva para la madre y el bebé, sino que este mismo se convierte en un riesgo de desarrollo de enfermedades para el infante prematuro. Las causas de este parto son de tipo multifactorial ya que influyen las características físicas de la madre como condiciones ambientales o sociales. Es por ello, que en este trabajo se plantea el desarrollo e implementación de un software basado en aprendizaje automático para poder estudiar cómo diversas variables de diferente índole pueden ayudarnos a predecir si en un embarazo hay riesgo de que ocurra un parto prematuro o no. En este trabajo, se analiza el conjunto de datos Mother's Significant Features mediante técnicas de análisis de datos y aprendizaje obteniendo como resultado un modelo de predicción con 10 variables que describen tanto aspectos físicos de la madre como aspectos socio-económicos. Este modelo se asocia a una aplicación web que mediante un cuestionario devuelve el valor de predicción del modelo. Con el desarrollo de este trabajo se ha demostrado que los factores que afectan al parto prematuro son de diferentes tipos (físico madre, estado salud o ambientales). Además, se ha facilitado una herramienta para la detección y seguimiento de embarazos de riesgo.</p>	

Abstract

Premature births are a problem of great importance that affects all countries of the world. Thus, of the 15 million births produced per year, 84% are considered premature. Not only do they imply a problem in terms of complications in childbirth and the danger that it entails for the mother and the baby, but it also becomes a risk of developing diseases for the premature infant. The causes of this delivery are multifactorial since the physical characteristics of the mother influence as well as environmental or social conditions. For this reason, this paper proposes the development and implementation of software based on automatic learning to be able to study how various variables of different kinds can help us predict whether or not there is a risk of premature birth in a pregnancy. In this paper, the Mother's Significant Features data set is analyzed using data analysis and learning techniques, obtaining as a result a prediction model with 10 variables that describe both physical aspects of the mother and socio-economic aspects. This model is associated with a web application that, through a questionnaire, returns the prediction value of the model. With the development of this work, it has been shown that the factors that affect premature birth are of different types (mother's physical condition, state of health or environmental). In addition, a tool for the detection and monitoring of risk pregnancies has been provided.

ÍNDICE GENERAL

1. Introducción	7
1.1. Contexto y justificación del trabajo	7
1.2. Objetivos del trabajo	8
1.3. Impacto en sostenibilidad, ético-social y de diversidad	8
1.4. Enfoque y método seguido	10
1.5. Planificación del trabajo	14
1.6. Breve resumen de productos obtenidos	19
1.7. Breve descripción de los otros capítulos de la memoria	19
2. Estado del arte	20
2.1. Partos prematuros	20
2.2. Teoría sobre metodología	27
3. Materiales y métodos	40
3.1. Análisis de datos	40
3.2. Aprendizaje automático	44
3.3. Productos	47
4. Resultados	49
4.1. Análisis de datos	49
4.2. Aprendizaje automático - Definición modelos	54
4.3. Aprendizaje automático - Optimización modelos	58
4.4. Aplicación web	60
5. Conclusiones y trabajos futuros	62
5.1. Conclusiones	62
5.2. Seguimiento de la planificación	63
5.3. Impacto ético-sociales, sostenibilidad y diversidad	64
5.4. Líneas de futuro	65
6. Glosario	67
7. Bibliografía	68

Lista de figuras

Figura 1.1 Esquema de los tipos de aprendizaje automático	pág 12
Figura 1.2 Análisis exploratorio	pág 12
Figura 1.3 Planificación temporal PEC1	pág 17
Figura 1.4 Planificación temporal PEC2	pág 17
Figura 1.5 Planificación temporal PEC3	pág 18
Figura 1.6 Planificación temporal PEC4	pág 18
Figura 2.1 Datos informativos parto prematuro	pág 22
Figura 2.2 Causas parto prematuro	pág 26
Figura 2.3 Árboles de decisión	pág 33
Figura 2.4 Bosques aleatorios	pág 34
Figura 2.5 Máquinas de vector de soporte	pág 35
Figura 2.6 Red neuronal artificial	pág 37
Figura 2.7. Curva ROC y Área bajo la curva	pág 39
Figura 3.1 Análisis de datos	pág 43
Figura 4.1 Información del <i>dataframe</i> original	pág 49
Figura 4.2 Frecuencia tipo de partos	pág 50
Figura 4.3 Matrices de confusión en los métodos de detección de valores atípicos ..	pág 52
Figura 4.4 Conjunto de tablas de relevancia de variables	pág 53
Figura 4.5 Comparación código html y página web	pág 61
Figura 4.6 Resultado página web	pág 61

Lista de tablas

Tabla 1.1 Tareas PEC1	pág 15
Tabla 1.2 Tareas PEC2	pág 15
Tabla 1.3 Tareas PEC3	pág 16
Tabla 1.4 Tareas PEC4	pág 16
Tabla 4.1 Valores f1 y AUC	pág 54
Tabla 4.2 Valores f1 y AUC	pág 55
Tabla 4.3 Valores hiperparámetros y medidas de evaluación	pág 56
Tabla 4.4 Valores f1 y AUC	pág 56
Tabla 4.5 Valores <i>Loss</i>	pág 57
Tabla 4.6 Valores f1 y AUC	pág 58
Tabla 4.7 Valores f1 y AUC	pág 58
Tabla 4.8 Valores f1 y AUC	pág 59
Tabla 4.9 Valores f1 final de todos los modelos	pág 60

1. INTRODUCCIÓN

1.1. Contexto y justificación del trabajo

De acuerdo con la definición de la Organización Mundial de la Salud (OMS)¹, se considera un parto prematuro aquel que ocurre antes de la semana 37 de embarazo. Dentro del parto prematuro, podemos considerar cuatro tipos: extremadamente prematuro (menos de 28 semanas), muy prematuro (entre 28 y 32 semanas), moderado (entre 32 y 34 semanas) y prematuro “tardío” (34 a 37 semanas).

Aunque no se pueda conocer con toda certeza la incidencia del parto prematuro debido al difícil acceso a los datos en ciertos países, los últimos estudios realizados en 184 países estiman que de 15 millones de nacimientos que se producen al año, un 84% ocurren entre las semanas 32 y 36 de la gestación². No sólo el problema viene en el momento del parto y sus complicaciones, sino que además el nacimiento prematuro es la primera causa de muerte infantil y también conlleva problema de morbilidad a los infantes supervivientes. De hecho, de estos 15 millones de niños nacidos de manera prematura, se estima que más de un millón muere antes de los 5 años debido al parto prematuro y sus complicaciones³. La mayoría de las muertes ocurren en la región subsahariana de África y en la región sur de Asia. Por ejemplo, en India se producen el 33% de las muertes totales⁴.

Por otro lado, este problema también acontece a una parte de índole económico. En 2005, el Instituto de Medicina (IOM) de Estados Unidos calculó que en este país, los partos prematuros suponen un gasto de 26 millones de dólares al año; teniendo en cuenta los gastos del parto, la intervención y el cuidado de los niños hasta los 5 años⁵. Además, debido a la morbilidad de estos partos, el gasto médico incrementa ya que los infantes nacidos por parto prematuro tienen riesgo de desarrollar asma, problemas de aprendizaje, desorden de déficit de atención y problemas emocionales, entre otros⁶.

Debido a todos los problemas que conlleva a la sociedad este tipo de parto, en los últimos años muchos estudios se han centrado en identificar las causas de ellos estudiando tanto características médicas de las madres como también condiciones ambientales de la vida de ellas como puede ser tipo de vida o nivel socio-económico. Se ha visto que el parto prematuro ocurre por una serie de condiciones complejas que resultan de múltiples vías etiológicas como pueden ser si la madre es fumadora o no hasta el nivel económico de la madre⁷.

Debido a la complejidad del problema respecto al descubrimiento de las causas reales y la relación de estas en el parto, se puede plantear como posible solución el desarrollo e implementación de una de las herramientas que nos puede ofrecer el *Machine Learning*. Como consecuencia del crecimiento y desarrollo de esta tecnología en la última década, cada vez es más común el uso de diferentes algoritmos para funciones médicas como pueden ser la predicción de factores de riesgo en posibles pacientes de cáncer de pecho⁸ o en enfermedades inflamatorias de riñón en niños⁹. En nuestro caso, se plantea el desarrollo e implementación de un *software* basado en aprendizaje automático para poder estudiar cómo diversas variables de diferente índole (biomédicas y sociales) pueden ayudarnos a predecir si en un embarazo hay riesgo de que ocurra un parto prematuro o no. Así pues, se desarrolla este modelo que junto con la creación de una aplicación web permite conocer si una mujer tiene posibilidades de tener un parto prematuro o no.

1.2. Objetivos del trabajo

A. Objetivo general

- 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.

B. Objetivos específicos

Primer objetivo:

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

Segundo objetivo:

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

1.3. Impacto en sostenibilidad, ético-social y de diversidad

Cumpliendo con el compromiso público que mantiene la UOC con los objetivos de desarrollo sostenible 2030 de la ONU¹⁰ y afianzando la aplicación de los competencias éticas y global que se exigen en este trabajo; en este apartado, se procede a explicar la identificación de los impactos positivos y negativos que tiene el presente trabajo con respecto a las tres dimensiones de la competencia transversal de la UOC, siendo estos tres: sostenibilidad, comportamiento ético y responsabilidad social y diversidad y derechos humanos. Para una mayor claridad de estos aspectos y facilitando así el seguimiento de este mismo trabajo, procedemos a explicar cada apartado separando los impactos con respecto a las tres dimensiones anteriormente mencionadas.

A. Sostenibilidad

Según la ONU, podemos entender el concepto sostenibilidad como lo que permite *“satisfacer las necesidades del presente sin comprometer la habilidad de las futuras generaciones de satisfacer sus necesidades propias”*¹¹. En este caso, centramos esta definición en los aspectos referidos al impacto medioambiental y ecológico buscando el impacto de este presente trabajo en aspectos de la sostenibilidad medioambiental.

Sabiendo que este trabajo se centra en el estudio de variables que influyen en los partos prematuros, se ha podido estudiar si la calidad del aire es uno de los factores que afectan a los partos prematuros conociendo de antemano anteriores estudios que han postulado este factor como plausible¹². No obstante, aunque esta variable estuviese incluida en el conjunto de datos inicial y se haya mostrado que pueda afectar, en este trabajo no se ha considerado un objetivo principal el estudio del impacto de variables relacionadas con el medioambiente¹ (coincidiendo con el OD 11 - *Climate action*). Este criterio vino dado y limitado por la tipología de variables que han sido recogidas en el estudio inicial, siendo las implicadas en un desarrollo sostenible medioambiental prácticamente nulas.

Es por ello, que se considera que este trabajo no tiene ningún impacto directo a nivel de sostenibilidad medioambiental. Se podría considerar a modo de consecuencia secundaria, la reducción tanto de medios como de residuos si con los resultados de este trabajo se consiguiese evitar la cantidad de partos prematuros. Evitando así el uso de recursos sanitarios destinados a paliar las complicaciones y/o enfermedades que derivan de los partos prematuros.

B. Comportamiento ético y responsabilidad social

Con respecto a esta dimensión, se pretende describir el compromiso del trabajo a contribuir voluntariamente a una sociedad más justa usando la reflexión y aplicación de posibles formas de resolver los problemas con una responsabilidad personal y social con respecto a la sociedad. Esta responsabilidad se centrará en aspectos éticos-sociales como la pobreza, condiciones de vida y/o la paz, justicia e instituciones sólidas.

Para la realización de este trabajo se ha usado una base de datos con pacientes provenientes de India. Como se ha comentado anteriormente en esta memoria, en India ocurren un 1/5 de los partos prematuros a nivel mundial, destacando que este país alcanzó los máximos de partos prematuros en el 2020¹³. Uno de los motivos que se conocen que afectan a los partos prematuros es la condición social y económica de la madre¹⁴, siendo este también un factor que se ha tenido en cuenta en el conjunto de datos con el que se ha trabajado durante el desarrollo de este mismo trabajo. Por otra parte, cabe destacar que en los últimos estudios sobre desigualdades sociales y económicas a nivel mundial (realizados en 2022), India ha superado sus cifras ya que solamente al 10% de la población le corresponde el 57% de los ingresos del país¹⁵ contrastando con el 13% de ingresos que obtienen los estratos más pobres. Al menos un 30% de estas desigualdades vienen marcadas por casta, género o estrato social-económico de la familia.

Por todas estas razones, uno de los objetivos sociales a los que se puede comprometer este trabajo es a la reducción de las desigualdades dadas por el estrato social y económico en el seguimiento médico de las mujeres embarazadas (coincidiendo con el ODS1- *No poverty*). Con el desarrollo de la aplicación web se intenta construir un medio fácil y asequible que permita que en la mayoría de consultas médicas se puedan recoger los datos de las pacientes correspondientes a las variables finales definidas por el modelo. Por tanto, aportar la predicción de los partos prematuros a cualquier consulta y/o hospital permitirá poner en conocimiento el estado del embarazo a las pacientes y además aportar la opción de un seguimiento más exhaustivo. Con este fin, se intenta reducir el porcentaje de partos prematuros en mujeres embarazadas en general, evitando que la condición social y económica sea un factor influyente.

C. Diversidad y derechos humanos

Entendiendo la diversidad como el conjunto de personas con diferentes características físicas, sociales y personales en un grupo¹⁶ y sabiendo que los derechos humanos son aquellos derechos que corresponden a cualquier persona por el simple hecho de su condición humana¹⁷; en esta última dimensión se estudia la implicación de este mismo trabajo a estos aspectos de índole social pudiendo comprobar si se genera algún impacto con respecto a estos dos puntos.

Habiendo explicado anteriormente el impacto de este trabajo en el seguimiento de embarazos con riesgo de partos prematuros independientemente de la clase social, se puede afirmar que los efectos positivos de este trabajo velan por el cumplimiento de los derechos humanos. En concreto podemos referirnos al artículo 25.2 que expresa: “*La maternidad y la infancia tienen derecho a cuidados y asistencia especiales.*”¹⁸, razonando así el desarrollo de la aplicación web que permita dotar de más recursos al personal sanitario para poder velar por el cuidado tanto de la madre durante el período gestante como del infante evitando su nacimiento prematuro y las posteriores complicaciones.

Por otro lado, podemos afirmar que este trabajo vela por una igualdad de género intentando prestar un servicio que mejore la asistencia sanitaria a las mujeres embarazadas, sabiendo de antemano que los estudios científicos suelen estar sesgados por género y no estar centrados en aspectos específicos de la salud de la mujer como puede ser la menstruación y/o embarazo¹⁹.

Todos los ejemplos expuestos demuestran que el desarrollo de este trabajo está firmemente comprometido tanto en el cumplimiento de los derechos humanos (centrándonos en los derechos de la madre y el infante) como con la perspectiva de género (ODS 5- *Gender equality*; ODS 10- *Reduced inequalities*). La decisión final sobre la elección de la base de datos para la realización de este trabajo vino influenciada por el deseo de querer aportar las herramientas más modernas como pueden ser los algoritmos de aprendizaje automático a temas de relevancia médica y social. Siendo así que con el desarrollo tanto de los modelos de aprendizaje automático como con la aplicación web, se podrá asegurar una atención con más calidad a las mujeres embarazadas. Además, este estudio aporta más detalles sobre los factores que influyen en los partos prematuros colaborando así en los diversos estudios realizados al respecto.

1.4. Enfoque y método seguido

El aprendizaje automático es una rama de la informática que nos permite descifrar patrones de cualquier conjunto de datos para mejorar el rendimiento en varias tareas²⁰. Con el desarrollo de los algoritmos y la implementación de estos en los problemas que acontecen a diferentes ámbitos como medicina, sector financiero, etc., el aprendizaje automático se ha convertido en una de las herramientas más necesitadas y útiles en los últimos tiempos.

En el caso del ámbito de la salud, el verdadero lugar del aprendizaje automático no es de la explicación de los datos, sino es el de la predicción; ya que en este caso, para poder explicar los datos se necesita una visión experta que entienda bien la casuística de estos. Es por ello que los algoritmos de aprendizaje automático modernos se orientan con el objetivo central de predecir de manera flexible los resultados de los nuevos datos con la mayor precisión posible²¹. Esto nos permite aliviar muchas de las suposiciones sólidas

detrás de los modelos clásicos, permitiendo que la conexión entre las covariables y un resultado sea mediada por cualquier algoritmo de caja negra, ahorrando consideraciones de interpretabilidad y plausibilidad para discusiones post-hoc.

En nuestro caso, para la realización de este trabajo partiremos de un conjunto de datos llamado “*Mother’s Significant Feature (MSF)*” extraído del repositorio *IEEE Data Port*²². Este conjunto de datos ha sido diseñado con el objetivo de proporcionar información a los investigadores que trabajan para mejorar la salud de la mujer y el niño. Los registros del conjunto de datos MSF han sido recopilados en la región metropolitana de Mumbai, específicamente en Maharashtra, India. Las mujeres fueron entrevistadas justo después del parto entre febrero de 2018 y marzo de 2021.

MSF consta de 450 observaciones individuales con un total de 130 variables que consisten en características de la madre, algunas características del padre y resultados de salud durante el embarazo. Se crea un conjunto de datos detallado para comprender las características de la madre que se pueden clasificar en 5 categorías: física, social, estilo de vida, nivel de estrés y resultados de salud. Además, también para algunas variables se diferencian los resultados en las tres fases de la edad reproductiva: adolescencia, después del matrimonio y durante el embarazo.

Como se ha explicado anteriormente y volviendo a nuestro tema interés, India es el país donde ocurren el 33% de las muertes a causa de un parto prematuro por tanto usamos este conjunto de datos para conocer qué factores influyen más en el parto prematuro en este país donde hay más incidencia. Además, dentro de este conjunto de datos tenemos la variable de parto prematuro la cual es nuestra **variable dependiente**. Es por ello, que el tipo de algoritmos que usaremos para la realización de este trabajo serán los concebidos dentro del **aprendizaje supervisado**, ya que conocemos de antemano la etiqueta de los tipos de parto. Este tipo de aprendizaje funciona de la forma que a medida que los datos van siendo introducidos en el modelo, este ajusta sus ponderaciones hasta que dicho modelo se haya ajustado adecuadamente, lo que ocurre como parte del proceso de validación cruzada²³.

Además, este tipo de aprendizaje se suele usar para dos tipos de problemas:

- Clasificación: Utiliza un algoritmo para asignar con precisión datos de prueba en categorías específicas. Reconoce entidades específicas dentro del conjunto de datos e intenta sacar algunas conclusiones sobre cómo esas entidades deben etiquetarse o definirse. Los algoritmos de clasificación comunes son clasificadores lineales, máquinas de vectores de soporte (SVM), árboles de decisión, k vecinos más próximos (*k-nearest neighbors*, *kNN*) y bosques aleatorios (*Random Forest*, *RF*).
- Regresión: Se utiliza para comprender la relación entre variables dependientes e independientes. La regresión lineal, la regresión logística y la regresión polinomial son populares algoritmos de regresión.

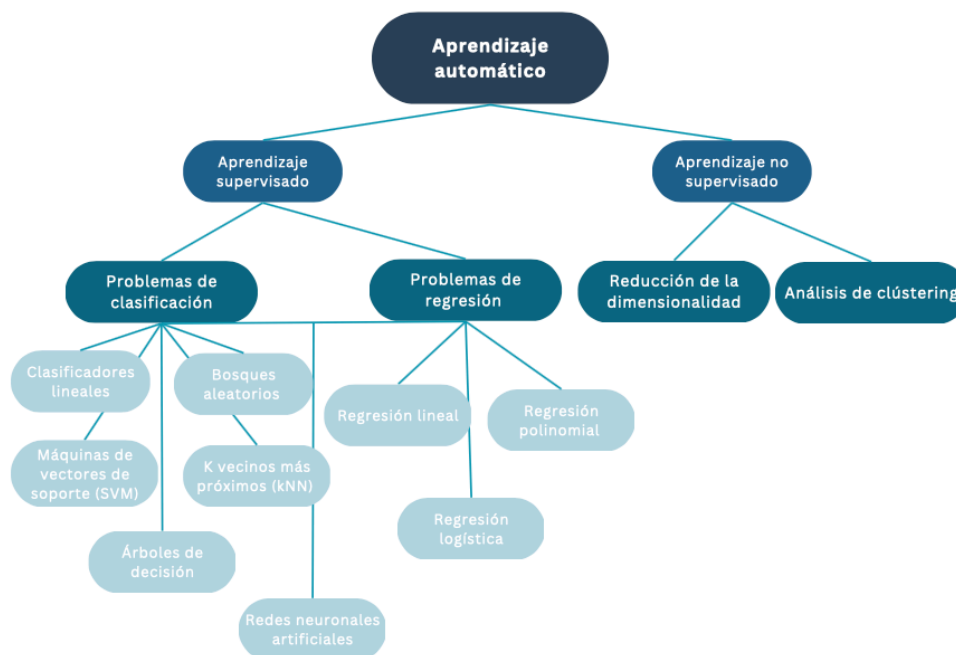


Figura 1.1. Esquema de los tipos de aprendizaje automático. Mapa conceptual del tipo de aprendizaje automático.

Como se puede apreciar en la imagen inferior, se parte de un conjunto de datos con un número considerable de variables. El primer objetivo es mejorar el manejo de este conjunto de datos realizando un paso previo antes de la definición de los diferentes modelos, intentando controlar qué variables son las más importantes con respecto a la **variable dependiente** (parto prematuro o no). Para ello, se usan técnicas de reducción de la dimensionalidad como *Random Forest Classifier*. También como se puede observar, se realiza el análisis y el posterior desarrollo de los modelos usando el lenguaje de programación Python.

```
datos = pd.read_excel('TFM_MSF_Dataset_Complete_450.xlsx')
```

```
datos.head()
```

	Mother_UID	Age_Of_Mother	weight_before_preg	wt_before_delivery	Height(cm)	BMI	Hemoglobin	PCOS	Age_Father	Yrs_Of_Marriage	...	Full Term	births
0	1	29	59	60.0	156	25	12.5	0	31	5	...	1	
1	2	24	54	56.0	145	26	12.5	0	28	2	...	1	
2	3	28	62	65.0	151	28	11.5	0	31	4	...	1	
3	4	25	49	52.0	151	22	11.5	0	30	3	...	1	
4	5	21	39	42.0	151	18	10.1	0	25	2	...	1	

5 rows x 131 columns

```
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
```

```
len(datos)
```

450

Figura 1.2. Análisis exploratorio. Análisis exploratorio del conjunto de datos “MSF” usando el lenguaje de programación Python.

Este apartado, no sólo ayuda a tener un posterior mejor manejo de los datos pudiendo así facilitar la aplicación de los modelos; sino que también, colabora a la definición de un nuevo fenotipo sobre los partos prematuros facilitando un nuevo sistema de clasificación y pudiendo dar un patrón para seguir en hospitales.

Por otra parte, teniendo ya el conjunto de datos con las variables más representativas; se procede a la aplicación de diferentes algoritmos dentro del aprendizaje supervisado centrándonos en los incluidos en el tipo clasificación. Como hemos mencionado, existen varios tipos de algoritmos pero como se trabaja con un conjunto de datos con diferentes variables que en su mayoría son categóricas, consideramos que los algoritmos que más se adaptan a nuestras necesidades son árboles de decisión (*Decision Tree, DT*), bosques aleatorios, máquinas de vectores de soporte y redes neuronales (*artificial neural network, ANN*).

La decisión de la elección de estos algoritmos se basa tanto en las características propias de estos como en el conjunto de datos²⁴:

- Árboles de decisión: Es uno de los primeros algoritmos que se desarrolló en el aprendizaje automático. Modeliza las pruebas de decisión como si fuese un test y los resultados los transforma a modo de estructura de árbol. Dependiendo del resultado de la prueba, el algoritmo se bifurca hacia un nodo secundario apropiado donde el proceso de prueba y bifurcación llega al nodo hoja.

Este algoritmo es fácil de usar, es compatible con múltiples tipos de datos (categóricos, nominales, etc) y genera clasificadores robustos que pueden ser validados usando test estadísticos.

- Bosques aleatorios: Es un clasificador de conjunto y consta de muchos árboles de decisión asemejándose a la creación de un “bosque” como conjunto de árboles. Los árboles de decisión de un bosque aleatorio se entrenan utilizando las diferentes partes del conjunto de datos de entrenamiento. Para clasificar una nueva muestra, se requiere que el vector de entrada de esa muestra se transmita con cada árbol del bosque. Luego, cada árbol considera una parte diferente de ese vector de entrada y da un resultado de clasificación. Luego, el bosque elige la clasificación de tener la mayor cantidad de “votos”.

Este algoritmo funciona bien para conjuntos de datos grandes, además ayuda a conocer cuáles son las variables más importantes en la clasificación.

- Máquinas de vectores de soporte (SVM): Es un algoritmo que se puede usar tanto para datos lineales como no lineales. Primero mapea cada elemento de datos en un espacio de características n -dimensional donde n es el número de características. Luego identifica el hiperplano que separa los elementos de datos en dos clases mientras maximiza la distancia marginal para ambas clases y minimiza los errores de clasificación.

Es más robusto que la regresión logística pudiendo manejar múltiples espacios y tiene menos riesgo de sobreajuste.

- Redes neuronales artificiales: Intentando asemejarse a la estructura de un cerebro, este algoritmo presenta un grupo de nodos interconectados. La salida de uno de los nodos va como entrada a otro nodo para su posterior procesamiento según la interconexión. Los nodos normalmente se agrupan en una matriz llamada capa dependiendo de la transformación que realicen. Además de la capa de entrada y salida, puede haber una o más capas ocultas.

Es una de las herramientas más potentes de aprendizaje profundo ya que se pueden detectar relaciones no lineales complejas entre las variables dependientes e independientes, permitiendo la adaptación a problemas de clasificación y regresión.

Para finalizar, en cada uno de los modelos seguiremos estos pasos básicos para trabajar con aprendizaje automático²⁵:

1. Recogida y exploración de los datos.
2. Creación de un conjunto de test y otro de entrenamiento.
3. Preprocesamiento de los datos.
4. Construcción de los modelos escogidos.
5. Entrenamiento y validación de los modelos.
6. Optimización de los modelos.
7. Validación del modelo más óptimo.

Con respecto al desarrollo de la aplicación web, existen multitud de opciones para la realización de este paso. El paso más común para las aplicaciones web es el uso de lenguajes conocidos en este entorno como *html*, *css* y *MySQL* pudiendo desarrollar la estructura y estética de una web y vincular esta misma a una base de datos conocida. No obstante, debido a que el resultado que se obtiene en este trabajo es un modelo de aprendizaje automático se han buscado alternativas que faciliten la implementación de este tipo de resultados. Se encuentra la opción del *microframework* de Flask que permite el desarrollo de la aplicación web usando lenguaje Python simplemente creando un nuevo cuaderno de Jupyter. Por tanto, el desarrollo del *BackEnd* correspondiente a la aplicación web se desarrolla usando Flask. La estructura y estilo se define usando *html* y *css*. Por último, para la conexión de estos archivos con un servidor, se usa el IDE (*integrated development environment*) de *PythonAnywhere*.

Para finalizar la creación de este trabajo, se publica todo el código implementado en este proyecto en un repositorio público como Github pudiendo así facilitar la corrección de este mismo trabajo.

1.5. Planificación del trabajo

A. Tareas

En este apartado, se explica de manera más detallada las diferentes tareas que se han realizado durante el desarrollo del trabajo de fin de máster. Para poder mejorar el seguimiento de estas tareas y así acotar de manera más clara los tiempos de entrega aproximado, las tareas correspondientes a cada objetivo se van a desglosar siguiendo las entregas pautadas por la propia asignatura. Para ello, se exponen cada grupo de tareas en una tabla correspondiente a cada PEC.

Para la primera PEC, nos basamos en un estudio bibliográfico tanto de los modelos de aprendizaje automático existentes centrándonos en los más utilizados en el ámbito de la salud. Además, también se investigó sobre el tema de interés para este trabajo (parto prematuro). Conociendo con más detalle estos dos pilares de nuestro trabajo, procedimos a la redacción del proyecto de memoria y la entrega.

PEC 1 - DEFINICIÓN Y PLAN DE TRABAJO (01/03 - 20/03)	
Descripción	Fecha inicio - final
Recopilación del conjunto de datos MSF y análisis exploratorio	01/03 - 03/03
Búsqueda bibliográfica y lectura sobre aprendizaje automático	04/03 - 06/03
Búsqueda bibliográfica y lectura sobre parto prematuro	06/03 - 08/03
Contextualización y descripción de la PEC1	08/03 - 10/03
Redacción de la PEC1 - Plan de trabajo	11/03 - 17/03
Revisión y entrega de la PEC1 - Plan de trabajo	18/03- 20/03

Tabla 1.1. Tareas PEC1. Tareas de manera desglosada correspondientes a la PEC 1.

Siguiendo las pautas marcadas por la asignatura, se abordó el primer tramo del desarrollo del trabajo en sí. Para ello, primero nos centramos en la parte de exploración y preprocesamiento de datos. Esto viene incluido el cumplimiento tanto del esquema general del trabajo en modelos de aprendizaje automático como de nuestro primer objetivo general descrito. En concreto, este primer tramo de desarrollo del trabajo se centró en cumplir el primer objetivo.

PEC 2 - DESARROLLO DEL TRABAJO: FASE 1 (21/03 - 24/04)	
Descripción	Fecha inicio - final
Estudio de las variables y descripción detallada del conjunto de datos	21/03 - 28/03
Estudio y aplicación de reducción de dimensionalidad para variables	29/03 - 06/04
Primer objetivo: Obtención de conjunto de datos óptimo	07/04
Adaptación del conjunto de datos para cada algoritmo y construcción de modelos	08/04 - 13/04
Entrenamiento de los modelos correspondientes	14/04 - 20/04
Redacción y entrega de PEC 2	21/04 - 24/04

Tabla 1.2. Tareas PEC2. Tareas de manera desglosada correspondientes a la PEC 2.

En esta tercera fase, se procedió a la revisión de los errores que se encontraron en la primera fase de desarrollo del trabajo (correspondiente a la PEC 2) como la revisión de los métodos de reducción de dimensionalidad. Luego, se acabaron de establecer todos los modelos de aprendizaje automático pertinentes y se realizó la validación de estos. Por último, se desarrolló la aplicación web finalizando así el contenido más “grosso” del trabajo de fin de máster.

PEC 3 - DESARROLLO DEL TRABAJO: FASE 2 (25/04 - 29/05)	
Descripción	Fecha inicio - final
Revisión y corrección de errores de la PEC 2	25/04 - 28/04
Aplicación de las mejores de los modelos y validación de las mejoras	29/04 - 07/05
Aprendizaje de desarrollo de aplicación web mediante Flask	08/05 - 12/05
Desarrollo de la aplicación web de Flask	12/05 - 20/05
Comprobación del funcionamiento y corrección de errores de la aplicación web	21/05 - 24/05
Segundo objetivo: Aplicación web con modelo óptimo para el conjunto de datos	23/05 - 24/05
Redacción y entrega de PEC 3	25/05 - 29/05

Tabla 1.3. Tareas PEC3. Tareas de manera desglosada correspondientes a la PEC 3.

En la última fase, se enmiendan los errores que no se pudieron resolver en la PEC3 como por ejemplo la optimización del código de las redes neuronales. Luego se procede a finalizar este trabajo con la publicación en un repositorio público del código junto con la redacción final del trabajo de fin de máster. Cabe destacar, que aunque en este tramo se especifique un apartado especial de “redacción del trabajo de fin de máster”, esta tarea se ha realizado durante los meses previos. Solamente se destaca y se deja un apartado específico en esta fase ya que al encontrarse en la parte final del proyecto, se ha considerado el tiempo específico a la revisión y corrección de la memoria en sí.

PEC 4 - CIERRE DE LA MEMORIA Y PRESENTACIÓN (30/05 - 20/06)	
Descripción	Fecha inicio - final
Corrección de errores de PEC 3	30/05 - 04/06
Optimización del código y publicación en repositorio público	05/06 - 06/06
Redacción del trabajo de fin de máster	07/06 - 12/06
Creación de la presentación y vídeo	13/06 - 19/06
Revisión y entrega de la PEC4	20/06

Tabla 1.4. Tareas PEC4. Tareas de manera desglosada correspondientes a la PEC 4.

B. Planificación temporal

Para la explicación de la planificación temporal seguida durante el desarrollo de este trabajo, se muestra la información detallada de todas las tareas realizadas en un diagrama de Gannt. Para poder visualizarlo de manera más clara, se crean diferentes diagramas correspondientes a cada PEC.

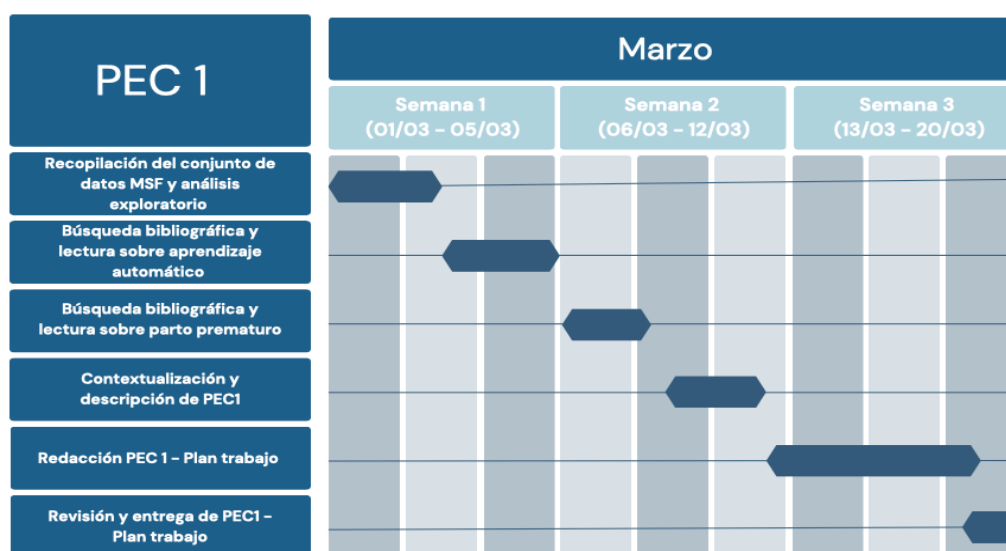


Figura 1.3. Planificación temporal PEC1. Diagrama de Gannt correspondiente a las tareas realizadas durante la PEC 1.

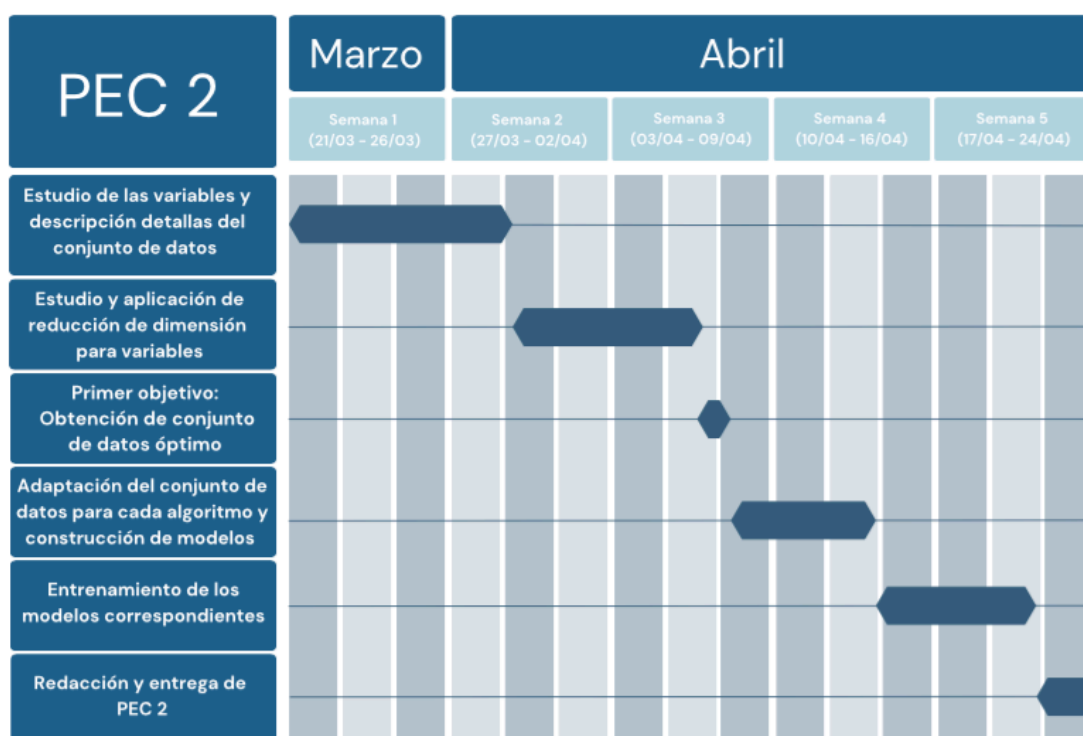


Figura 1.4. Planificación temporal PEC2. Diagrama de Gannt correspondiente a las tareas realizadas durante la PEC 2.

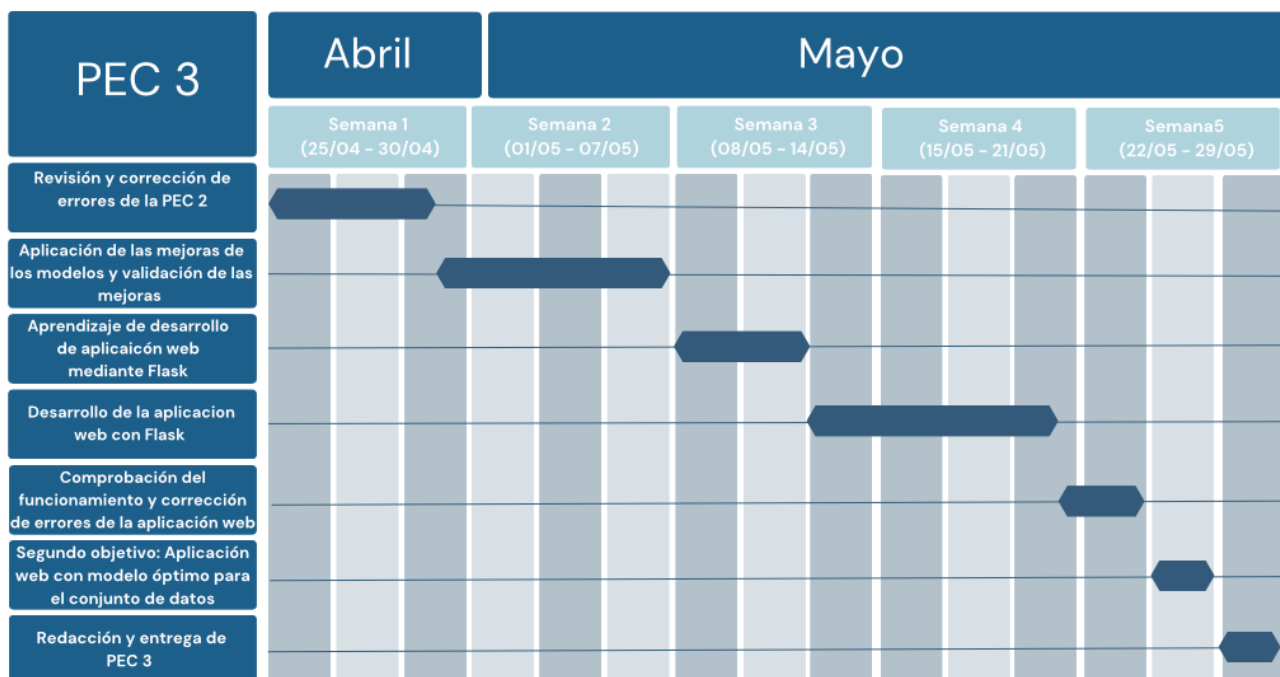


Figura 1.5. Planificación temporal PEC3. Diagrama de Gannt correspondiente a las tareas realizadas durante la PEC 3.

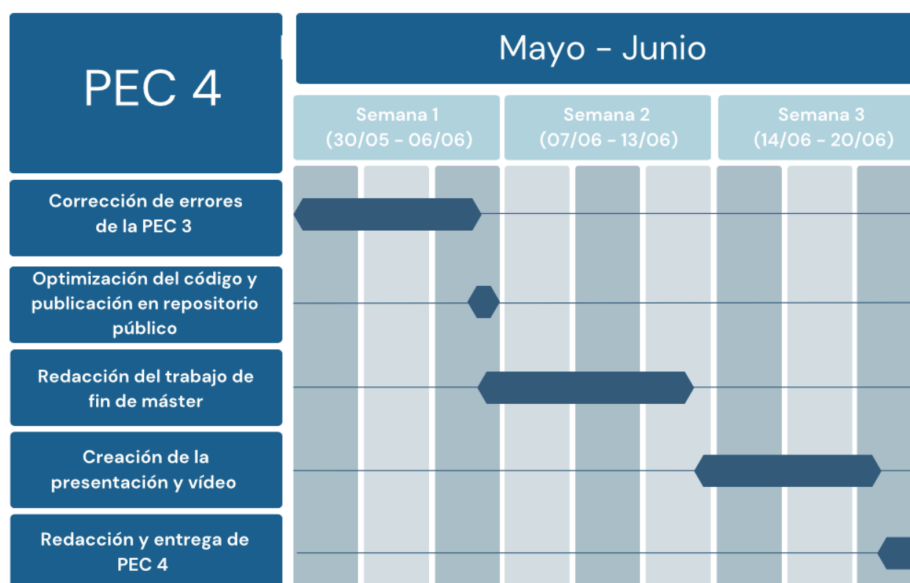


Figura 1.6. Planificación temporal PEC4. Diagrama de Gannt correspondiente a las tareas realizadas durante la PEC 4.

C. Recursos utilizados

El desarrollo de este trabajo se basa en gran parte en el uso del lenguaje de Python enfocado en el ámbito del análisis de datos y aprendizaje automático. Es así que los recursos utilizados en este trabajo han sido las documentaciones oficiales de las librerías más influyentes en este ámbito como son Tensorflow²⁶ y Keras²⁷, como también diferentes páginas webs de consulta sobre diferentes partes del código implementado^{28,29}.

Con respecto a la parte más teórica tanto de los algoritmos utilizados como de los partos prematuros, se han consultado libros teóricos y artículos de relevancia en cada uno de los temas. Estos recursos han sido referenciados en el apartado de “enfoque y método seguido”.

Como información más técnica, el entorno de desarrollo utilizado durante todo este trabajo ha sido cuaderno de Jupyter (versión 6.4.12), usando dentro de este la versión de Python de 3.9.13. Por otro lado, para el desarrollo web se ha usado tanto el cuaderno de Jupyter con la misma versión de Python, como el entorno de *PythonAnywhere* para el manejo del archivo html. Por último, todo el código implementado en el desarrollo de este proyecto se encuentra alojado en el repositorio público de *Github*.

1.6. Breve resumen de productos obtenidos

- **Aplicación web** desarrollada con el objetivo de la predicción del parto prematuro. Se facilitará un cuestionario que permitirá la obtención de las variables previamente establecidas como relevantes para cada paciente y así poder realizar la predicción. Se puede acceder a esta web mediante [este enlace](#).
- **Repositorio público** que nos permita acceder al código desarrollado durante todo el procedimiento de la creación de este propio trabajo. Se accede al repositorio por [este enlace](#).
- **Memoria escrita** del presente trabajo que incluye la base teórica y explicación detallada de los pasos realizados para la obtención de los dos productos anteriores.

1.7. Breve descripción de los otros capítulos de la memoria

Estado del arte

Este capítulo se centra en la descripción y contextualización de la situación de los partos prematuros a nivel mundial. Se hace hincapié en los datos referentes a la última década (2010-2020) viendo como la evolución de los partos prematuros no han mejorado durante este período de tiempo. Por otro lado, también se exponen las causas y efectos de los partos prematuros, centrando los datos en los niños. También se dedica una parte a explicar los efectos en las madres y las posibles soluciones que se barajan para arreglarlo, siendo una de ellas el uso del aprendizaje automático.

Por otro lado, se explican las bases teóricas de los pasos de análisis de datos y los algoritmos de aprendizaje automático dando al lector una base más sólida para poder entender las decisiones que se toman a lo largo de este trabajo.

Materiales y métodos

Se describe la metodología usada explicando cada paso de esta y justificando la elección de los métodos empleados. Se separa este apartado en tres grandes bloques: análisis de datos, aprendizaje automático y productos. En el análisis de datos, se explican los pasos de exploración de datos, valores nulos, valores atípicos y reducción de dimensionalidad obteniendo como paso final el conjunto de datos reducido y adaptado para la aplicación de modelos de aprendizaje automático. En el apartado de aprendizaje automático se

explican los pasos de implementación de los cuatro algoritmos escogidos (DT, RF, SVM y ANN) y las posteriores modificaciones en el caso de requerir una optimización de los modelos. Por último, en los productos se explica cómo se ha desarrollado la web usando Flask, html y *PythonAnywhere*.

Resultados

En este apartado se describen los resultados obtenidos a lo largo del desarrollo del proyecto. Aparece repartido en cuatro subapartados correspondientes a los resultados de análisis de datos, a los resultados de los modelos de aprendizaje automático, los resultados de los modelos después de la optimización y por último al resultado de la aplicación web. En el análisis de datos se aplican diferentes estrategias para el tratamiento de valores nulos dependiendo del tipo de variables, en los valores atípicos se decide no eliminar ninguno de los valores y, por último, se obtienen 3 conjuntos de datos con las 10 variables más importantes. Para los dos subapartados de aprendizaje automático, se aplican los algoritmos y se comparan los valores f1 y AUC teniendo que repetir la ejecución del modelo si no se supera el 75%. Por último, en la aplicación web se muestran la página web final.

Conclusiones y trabajos futuros

Este apartado final se divide en cuatro aspecto básicos que son las conclusiones finales obtenidas de la obtención y cumplimiento de cada uno de los objetivos del trabajo; se valora el desarrollo hecho desde un conjunto de datos muy extenso hasta la obtención de 10 variables siendo estas coincidentes en los tres conjuntos de datos que se barajan. El seguimiento de la planificación, exponiendo el cumplimiento de la línea temporal expuesta en los diagramas de Gannt y la adaptación en caso de no haber podido llevarlo todo a cabo. También se comenta el cumplimiento de los impactos predichos en el apartado 1.3, siendo en este caso un cumplimiento de todos los impactos positivos. Por último, se postulan líneas de trabajo futuras que permitan mejorar este estudio siendo una opción la ampliación del número de variables o la realización de un nuevo estudio con un enfoque centrado en nuevos aspectos como el efecto del cambio climático.

2. ESTADO DEL ARTE

En este capítulo de la memoria, se expondrá de manera detallada la importancia del trabajo justificándolo con la contextualización de éste en la sociedad, las hipótesis de trabajo y la metodología seguida finalizando con una explicación más detallada de los productos obtenidos. Para poder realizar un seguimiento congruente de este apartado, se propone la creación de tres subsecciones: introducción, metodología y productos.

2.1. Partos prematuros

A. Partos prematuros: ¿Problema de salud mundial?

Los partos prematuros y las complicaciones derivadas de ello se han postulado como uno de los grandes problemas de salud a nivel mundial. De hecho, Naciones Unidas cataloga este problema como una emergencia silenciosa, ya que cada año se cobra más de un

millón de vidas de los infantes prematuros. En la última década, los datos correspondientes a partos prematuros no han sufrido ningún cambio en ninguna región del mundo, siendo así que del 2010 al 2020 se calcula que han habido 152 millones de bebés vulnerables debido a su nacimiento prematuro³⁰.

Tal como indica el director de salud de UNICEF, Steven Lauwerie: *“A pesar de los avances en medicina que se han hecho en las últimas décadas, no se ha hecho ningún progreso en reducir el número de bebés prematuros o evitar el riesgo de muerte de estos. Es necesario mejorar el acceso a los cuidados tanto de las madres embarazadas como de los infantes prematuros y asegurar que cada niño tenga un inicio de vida saludable y prospere en la vida”*³¹.

Fijándonos en los datos, podemos corroborar la nula existencia de cambio en las tendencias del nacimiento prematuro, siendo un 9.9% del total de los partos en 2020 comparando esta cifra con el 9.8% de 2010. Esta carencia de cambio de tendencia también se muestra en regiones con un índice de partos prematuros más alto como son Asia Meridional con un 13.13% en 2010 a 13.2% en 2020 y África subsahariana que se mantiene con un 10.1% tanto en 2010 y 2020³².

A pesar de estos datos, también hay países, entre ellos España, que sí han reducido el porcentaje de partos prematuros en este período de 10 años. Por contra, otros países como Islandia o Chile lo han incrementado más del 5%. Este bagaje de datos implica que en esta década los partos prematuros han decrecido levemente, pasando de 13.8 millones a 13.4 millones debido a la disminución de partos. No obstante, estos datos siguen suponiendo un gran porcentaje de nacimientos prematuros, haciendo entrever la problemática de estos a nivel de salud global³².

La distribución de los partos prematuros a nivel mundial varía dependiendo de la zona y el país. Los datos muestran que la mayor incidencia de partos prematuros se encuentra en el sur de Asia, de hecho el 45% de los partos prematuros ocurridos en 2020 se concentran en 5 países: India, Pakistán, Nigeria, China y Etiopía. Como excepción, destacamos India por ser el país con más partos prematuros en 2020 (3.02 millones de nacimientos fueron prematuros, suponiendo estos un 23% del total a nivel mundial)³².

Con todos estos datos, se puede observar que existe una distribución desigual de los casos de nacimientos prematuros. Es por ello, que se debe prestar atención a los factores y causas de este problema. Los nuevos informes tanto de la OMS como de las NU (Naciones Unidas)³³ ponen en el foco cuatro problemas principales (referidos como las cuatro Cs, en inglés): conflicto, cambio climático, COVID-19 y la crisis del coste de vida.

- Conflicto: Se estima que más de 100 millones de personas en todo el mundo han emigrado debido a causas bélicas, siendo las mujeres y niños los más afectados^{34,35}. No sólo el estrés de la migración y las muertes directas cuentan, sino que las consecuencias como el colapso de sistema de salud o acceso restringido a este también influyen en estos datos. A nivel mundial, el 61% de las muertes de madres, el 51% de los niños muertos antes de nacer y el 50% de las muertes de recién nacidos ocurren en países que pidieron ayuda humanitaria a la ONU en 2023.
- Cambio climático: Un informe global realizado en 2020 estima que la contaminación del aire es la causa del 20% de las muertes de recién nacidos, muchos como resultado de un parto prematuro y/o un bajo peso al nacer³⁶. Además, el calor extremo cada vez más frecuente, se asocia con resultados adversos en el parto, incluyendo parto prematuro y

la muerte fetal^{37,38}.

- **COVID-19:** La pandemia del COVID-19 desestabilizó los servicios de salud tanto para las madres como para los recién nacidos. De hecho, se estima que debido a las restricciones en hospitales respecto a la madre y el niño, 125.000 muertes se podrían haber evitado si se hubiese permitido el contacto y cuidado de la madre con respecto al recién nacido³⁹.
- **Crisis coste de vida:** Debido a las interrupciones de suministros por el COVID-19, los conflictos bélicos y la crisis climática, existe una inflación a nivel mundial que ha aumentado del 4,7 a 8,8% en los últimos años⁴⁰. Los efectos con respecto a los bebés prematuros se han visto en informes donde se producen ingresos de bebés en cuidados intensivos debido a que la familia no puede pagar la calefacción⁴¹.

A pesar de ser estos 4 factores los más estudiados en los informes de estas dos organizaciones, se pueden extrapolar otras causas relacionadas como la situación socioeconómica. De hecho, estudiando los datos de incidencia de partos prematuros en continentes o países concretos, se puede observar que dentro de éstos, estas cifras varían. Destaca el caso de América Latina donde encontramos un 5.8% de nacimientos prematuros en Nicaragua, pero un 12.8% en Surinam³². Además, como se ha podido comprobar en este informe la incidencia más alta de partos prematuros se dan en países con ingresos bajos y medios, como son Bangladesh o Malawi. Estos datos muestran que la problemática de los partos prematuros está determinada por diferentes factores que no solo son los correspondientes a características médicas de la madre o complicaciones en el embarazo. Estas complicaciones también vienen seriamente marcadas por raza, etnia, nivel económico y acceso a un sistema de salud de calidad.

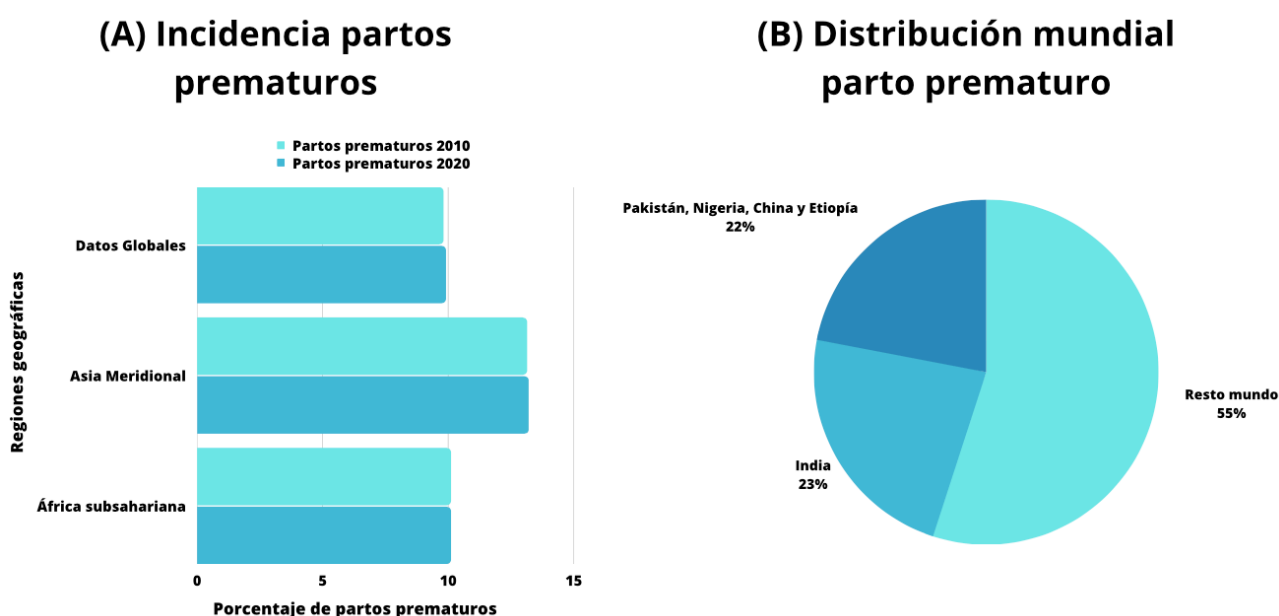


Figura 2.1 Datos informativos parto prematuro. (A) Se muestra un gráfico de barras con la incidencia a nivel mundial, en la región Asia Meridional y África subsahariana durante la década de 2010 a 2020. (B) Gráfico con la distribución mundial de partos prematuros en el 2020. Datos extraídos del informe de la ONU y Naciones Unidas⁴⁷.

Siendo ya conscientes de los factores de este problema y habiendo abordado éstos, procedemos a centrar la atención en los efectos de los nacimientos prematuros a nivel de salud global. Los partos prematuros constituyen la cuarta causa principal de pérdida de capital humano en todo el mundo. De hecho, las condiciones neonatales en general, incluidas la encefalopatía neonatal, las infecciones y condiciones congénitas, siguieron siendo la principal causa de la esperanza de vida reducida debido a discapacidad⁴². Además, las complicaciones en nacimientos prematuros son la primera causa de mortalidad infantil. Los datos recientes (recogidos en 2021) muestran que casi un millón de niños murieron a causa de complicaciones directas debido a los nacimientos prematuros, y sobre un tercio de los 2.3 millones de muertes prematuras son como consecuencia de las complicaciones en el parto prematuro⁴³.

No sólo es una causa directa de mortalidad infantil, sino que se asocia a perjuicios a largo plazo en los sistemas respiratorio y cardíaco, afectando también al desarrollo neurológico de los supervivientes. Estas afecciones pueden derivar en casos más graves como es la diplegia. Además, nuevos estudios muestran que los infantes nacidos incluso pocas semanas antes pueden desarrollar desordenes en el aprendizaje y comportamiento (siendo el caso de los nacidos entre la semana 32 y 37). De hecho, se ha visto que algunos nacidos entre la semana 37 y 39 tienen un riesgo elevado de resultados adversos en el desarrollo neurológico.

A pesar de todos estos datos, cabe remarcar que la mayoría de estos efectos pueden ser prevenidos mostrando así la calidad de la atención médica obtenida. Es por ello, que la problemática con los partos prematuros no afecta solamente a los mismos infantes, sino que también se convierten en un problema a nivel de cuidadores, sistema de salud y sociedad en general.

En todo este desarrollo, hemos podido ser conscientes del impacto que tienen los nacimientos prematuros en diversos aspectos de nuestra sociedad, poniendo siempre la atención en la afección directa a los niños (tanto nacidos o no). No obstante, este problema también afecta a las madres, siendo una de las partes implicadas en el proceso. Podemos ver que de 4.5 millones de muertes al año debido a los partos prematuros, 287.000 corresponden al fallecimiento de las madres, siendo los otros datos correspondientes a infantes que fallecen después del parto o a las muertes antes de nacer⁴⁴. El estudio y seguimiento de las mujeres embarazadas durante todo su proceso mediante unos servicios respetuosos y de alta calidad se plantean cómo fundamentales para mejorar los resultados de salud, incluidos la prevención y atención del parto prematuro.

Es cierto, que existen programas específicos (consensuados entre la OMS, UNICEF y UNFPA) que se centran en mejorar la calidad de atención sanitaria de las mujeres embarazadas y en empoderarlas. Este empoderamiento se basa en mejorar el acceso a servicios vitales en todo el espectro de salud sexual y reproductiva antes, durante y después del parto. A pesar del éxito de estos programas consiguiendo reducir las muertes maternas de manera significativa entre 2000 y 2015, cabe mencionar que este decrecimiento se estancó en 133 países, incluso se produjo un aumento de muertes en 17 países (pertenecientes mayoritariamente a Europa y América).

Uno de los factores que más afecta con respecto a la salud de las mujeres y los partos prematuros son la cobertura y la calidad del servicio sanitario. En este caso, la condición socio económica del país es uno de los factores más determinantes. En concreto, en países de América del Norte y Europa, la cobertura de atención prenatal llega a valores del 98%, en comparación con el 54% en África subsahariana. Además, dentro de cada

país se pueden ver desigualdades viendo la influencia del nivel económico, raza o etnia en la atención del embarazo⁴⁵. Uno de los ejemplos más llamativos es el caso de Kenia, donde los nacimientos atendidos por personal de salud calificado varían desde el 35% hasta el 92%.

A pesar que esta parte del estudio esté más centrada en las mujeres y el desarrollo del embarazo, los factores influyentes coinciden con los anteriormente mencionados (las cuatro Cs). De hecho, conflictos bélicos, enfermedades como el COVID-19 y el cambio climático son limitantes sobre el acceso a los servicios de salud de alta calidad. Siendo estos servicios esenciales para la prevención y manejo del parto prematuro. Como detalle, los datos de nueve países que se enfrentan a crisis humanitarias muestran que las tasas de mortalidad materna fueron más del doble del promedio mundial⁴⁶.

No obstante, en otros estudios se han barajado otros factores que puedan afectar siendo estos en su mayoría, características de la salud de la madre⁴⁷. Entre ellos encontramos:

- Edad de la madre,
- Tiempo entre partos,
- Embarazo múltiple,
- Infecciones,
- Enfermedades crónicas,
- Nutrición,
- Estilo de vida y tipo de trabajo,
- Efecto del ambiente,
- Salud mental,
- Desordenes genéticos.

Como podemos ver, aunque existan factores claramente influyentes como el caso de las 4 Cs, también hay otros que son de más fácil manejo como el estado de salud de la madre. Estos atributos permiten el diseño de diferentes estrategias para la confrontación e intento de solución de este problema.

B. Partos prematuros: ¿existen soluciones?

Durante el desarrollo del apartado anterior, se ha visto que la causa de los partos prematuros se podría considerar multifactorial ya que afectan tanto agentes externos de calidad social como pueden ser las crisis humanitarias, como también de calidad ambiente como el cambio climático. Por otro lado, encontramos que las características físicas de la madre (como nutrición o salud mental) son factores que nos pueden ayudar a entender y prevenir este tipo de parto.

Teniendo estos factores dependientes directamente de la salud y vida de la madre, se podrían plantear algunas medidas preventivas con respecto a los partos prematuros fijándonos en los factores más fácilmente controlables.

Ya existen algunas medidas que ayudan a mejorar o acelerar el desarrollo de los bebés que van a nacer prematuros, evitando así algunos de los problemas que puedan comprometer la supervivencia de estos. Un ejemplo, es la administración de corticosteroides de manera prenatal las cuales permiten acelerar el desarrollo de los pulmones⁴⁸. Esta medida facilita la respiración en los bebés prematuros, evitando cualquier complicación de tipo respiratorio.

Otra medida es la administración de medicamentos tocolíticos que pueden retrasar o incluso detener el parto prematuro, dando así más margen al funcionamiento de los corticosteroides y más tiempo para un traslado de la madre a un centro especializado. Además, se recomienda administrar sulfato de magnesio a las mujeres que probablemente den a luz antes de la semana 32, previniendo así la parálisis cerebral del bebé^{49,50}. En el caso en que se haya producido una rotura de membrana, la administración del antibiótico eritromicina puede prolongar el embarazo y prevenir morbilidades del infante. Las administraciones de los anteriores medicamentos se ven supeditadas al país que estudiemos, siendo una práctica extendida en países de ingresos altos, pero no tan usual en países de ingresos medios y bajos^{51,52}.

A pesar de seguir existiendo diferencias socio-económicas en las medidas de paliación de este tipo de nacimientos, la existencia de estos recursos abren una nueva puerta hacia posibles soluciones en los partos prematuros. De hecho, una de las opciones para intentar abordar este problema sería el estudio de métodos para la prevención de este tipo de nacimientos. En el caso de partos prematuros, la OMS describe tres tipos de prevención⁴⁷:

- Prevención primaria: Intervenciones realizadas directamente en las mujeres durante el embarazo.
- Prevención secundaria: Intervenciones dirigidas a mujeres con riesgo reconocido de parto prematuro. Esta detección incluye revisión clínica, medidas específicas como longitud del cervix o test fetal de fibronectina.
- Prevención terciaria: Intervenciones que se realizan después del inicio del parto prematuro, destinadas a mejorar la salud de los recién nacidos en los primeros años de vida.

Una de las formas en las que se puede abordar este problema es centrandose en nuevas investigaciones en la prevención primaria. Conociendo la existencia de factores que son fácilmente medibles con respecto a la salud y vida de la madre, se podría establecer un seguimiento del embarazo intentando predecir las probabilidades de cada madre de dar a luz de manera prematura. Esta solución podría evitar los problemas futuros que conllevan este tipo de nacimiento anteriormente explicados.

Varios estudios se han centrado en intentar conocer y establecer los verdaderos riesgos del parto prematuro durante el embarazo evitando centrarse en características físicas de la madre como la longitud del cérvix. Estas investigaciones se centran más en obtener información sobre diferentes aspectos de la vida de la madre como abusos de drogas o alcohol, estrato socioeconómico o factores demográficos, pudiendo así tener una visión más completa del perfil multifactorial de las mujeres gestantes^{53,54}.

Otro factor que contribuye tanto a la definición del perfil de madres con riesgo a tener un parto prematuro como al seguimiento del embarazo es la digitalización del sistema de salud. Mozambique es un ejemplo de país que a pesar de no tener una gran amplitud de recursos, ha logrado mejoras en su sistema de información de salud mediante el desarrollo de un sistema nacional de monitoreo y evaluación basado en una web desarrollada por el Ministerio de Salud y otras asociaciones. Esta web también se usa para la generación de datos nacionales sobre nacimientos prematuros y características de los recién nacidos⁵⁵. Este ejemplo nos muestra que la obtención de datos es uno de los

pasos cruciales para poder abordar el problema de los partos prematuros, no siendo necesaria una gran inversión económica. Además, la implementación de la tecnología en los sectores de salud, abre una vía sencilla y asequible para la creación de sistemas de seguimiento tanto de las madres como de los bebés nacidos de manera prematura⁵⁶.

Es por ello, que en este presente trabajo se plantea el uso de las nuevas tecnologías como son los algoritmos de aprendizaje automático para la predicción de parto prematuro en mujeres embarazadas. Este estudio se basará en el conjunto de datos *Mother's Significant Feature (MSF)* publicado en el repositorio público de *IEEE Data Port*²². Este conjunto de datos contiene 130 variables de carácter multifactorial ya que abarcan variables de tipo físico, social, estilo de vida, nivel de estrés y características de salud. Estas variables nos permiten realizar un estudio que abarca todos los factores que influyen en el parto prematuro expuestos en este mismo apartado centrándonos en la salud de la mujer y pudiendo así abordar el tipo de prevención primario que permite ir reduciendo los partos prematuros desde el origen.

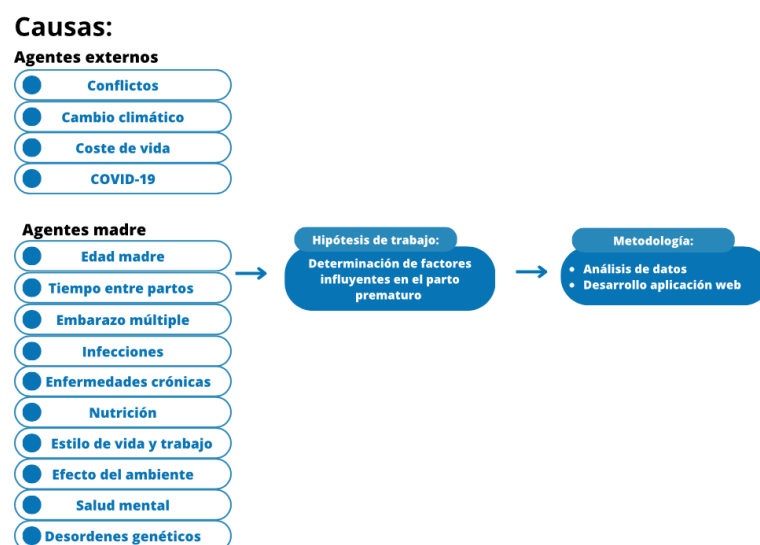


Figura 2.2 Causas parto prematuro. Esquema de causas parto prematuro, hipótesis y metodología del trabajo.

Es así que nuestra hipótesis de trabajo viene definida por la determinación de los factores más influyentes en mujeres embarazadas con respecto al riesgo de tener nacimientos prematuros. De esta forma, este trabajo pretende influir en las posibles soluciones para la disminución de este tipo de nacimientos. Esta investigación es un reflejo de la combinación de las últimas investigaciones que marcan la importancia de un estudio más general con respecto a las variables y la combinación de estos datos con técnicas tecnológicas que agilicen y aseguren este procedimiento. Además, no sólo se aplica el estudio de datos y los modelos de predicción, sino que siguiendo el ejemplo de Mozambique, se desarrolla una aplicación web para ofrecer un medio donde los diferentes hospitales o servicios sanitarios puedan actualizar estos datos. Esta aplicación permite no sólo la obtención de resultados directos para los casos que pasen por consulta, sino también se actualiza el modelo pudiendo reforzarlo cada vez más y mejorando su calidad con respecto a la capacidad de predicción.

2.2. TEORÍA SOBRE METODOLOGÍA

Debido a la complejidad del desarrollo de este trabajo y a la cantidad de técnicas implementadas durante éste, se considera oportuno añadir un subapartado que introduzca las bases teóricas de la metodología implementada. El objetivo de este apartado es ayudar al lector a la comprensión de cada uno de los pasos seguidos durante este mismo trabajo. Así, se puede entender la argumentación a favor de la metodología escogida posteriormente. Se distribuye este apartado en las fases:

- Implementación de técnicas de análisis de datos con el fin de poder obtener un conjunto de datos final con un número de variables manejable para la construcción de la aplicación web y su uso en servicios médicos.
- Uso de algoritmos de aprendizaje automático para poder obtener el modelo de predicción más adecuado a nuestros datos.

Para poder seguir de manera más clara la metodología utilizada en el desarrollo de este mismo trabajo, se van a crear subsecciones correspondientes a las fases anteriormente descritas. Dentro de las subsecciones solamente se explicarán aquellos pasos donde se necesita un conocimiento teórico de las diferentes opciones.

A. Análisis de datos

Preprocesamiento: Estudio de valores nulos

Se entiende como **valores nulos** aquellos valores donde no se ha almacenado ningún tipo de dato o no se tiene constancia de ello⁵⁷. Las causas de estos valores pueden ser desde pérdida de datos en sí, errores de entrada de datos o archivos faltantes. Estos valores suelen estar representados por la palabra “NaN” en el *dataframe*, siglas que significan “Not a Number”. Dentro de los valores nulos, podemos encontrar tres tipos:

- Valores NA completamente debidos al azar (MCAR, siglas en inglés): No existe una relación directa entre los datos faltantes y cualquier otro valor observado o no observado dentro del conjunto de datos. Este tipo de valor nulo se suele dar por un error humano, algún fallo del sistema o una pérdida de muestra.
- Valores NA parcialmente debidos al azar (MAR, siglas en inglés): El motivo por el que faltan valores puede explicarse por variables sobre las que sí que se tiene información completa ya que existe alguna relación entre los datos faltantes y otros valores. Se puede observar que faltan datos solo para algunas observaciones.
- Valores NA no debidos al azar (MNAR en inglés): Los valores nulos dependen de los datos no observados. Este tipo de valores nulos se da cuando existe una estructura en los datos faltantes y otros datos observados los pueden explicar. Se suelen dar debido a la reticencia de las personas encuestadas de responder algunas preguntas en la encuesta.

Para el tratamiento de valores nulos, existen dos soluciones para los valores nulos, por un lado está la opción de la eliminación de los valores nulos directamente evitando así todos

los problemas que puedan generar estos datos. Las opciones que existen para la eliminación son:

- Eliminación de la fila que contenga algún valor nulo: Si se da una fila con varios valores nulos, se puede eliminar esta observación directamente.
- Eliminación de la columna que contenga algún valor nulo: Si una columna contiene valores nulos, se puede eliminar la variable entera.

A pesar de que la eliminación de la fila o columna entera es la manera más sencilla de realizar el procesamiento de valores nulos, conlleva también la pérdida de información del conjunto de datos pudiendo hacer que se pierdan variables significativas que afecten al modelo que posteriormente se pueda implementar o que se pierdan muchas observaciones y por ello, las conclusiones que se puedan obtener de los modelos no sean estadísticamente significativas.

Por otro lado, existen diferentes técnicas de imputación que consisten en el reemplazo de estos valores nulos por unos que encajen en nuestro conjunto de datos⁵⁸. Dependiendo del tipo de dato al que se deba realizar este método, se pueden aplicar diferentes soluciones:

- Reemplazo por el valor de la media o mediana: Esta opción se usa para las variables numéricas de tipo continua. Se sustituyen los valores nulos por el valor de la media o mediana específica de esa variable.
- Reemplazo por valores “aleatorios” o nueva clase: Es válida tanto para variables numéricas como categóricas. Implica la creación de una nueva clase, sustituyendo todos los valores nulos por un número o clase que no exista previamente en esta variable. Permite la identificación de estos valores en pasos posteriores y la correcta evaluación del método de imputación.
- Reemplazo por valor común o moda: Se considera una opción exclusiva para datos de tipo categóricos o numéricos discontinuos. Se sustituyen los valores nulos por la categoría más común en la variable.
- Reemplazo por valor previo - valor siguiente: En el caso de tener variables de tipo series, se puede sustituir los valores nulos por el valor previo siguiendo la técnica de “*forward fill*” o por el valor siguiente basándose en la técnica de “*backward fill*”.
- Interpolación: En el caso de variables numéricas que sigan una distribución lineal, polinómica o cuadrática, se puede realizar una interpolación de los datos. Consiste en sustituir los valores nulos por los valores con más probabilidad de seguir la distribución de estos mismos datos.
- Uso de algoritmos para predicción de datos nulos: Se pueden realizar predicciones de los valores nulos usando algoritmos como kNN-vecinos o redes neuronales. La imputación realizada por kNN-vecinos se basa en la similitud de características para predecir los valores de los valores nulos. Al “nuevo punto” se le asigna un valor en función de su similitud con los puntos del conjunto de entrenamiento.

Dependiendo de las características de las variables que deben ser modificadas, se deberá escoger el tipo de tratamiento que se considere oportuno. No obstante, se recomienda el

uso de varias estrategias en el mismo conjunto de datos para poder luego validar qué método funciona mejor para el conjunto de datos con el que se trabaje⁵⁹.

Preprocesamiento: Estudio de valores atípicos

En estadística, se describen los **valores atípicos** como puntos de datos que no pertenecen a una población. Es decir, se trata de observaciones que se encuentran lejos o muy lejos de otros valores, siendo observaciones que divergen de los datos bien estructurados.

Dentro de los pasos de preprocesamiento en aprendizaje automático, el paso de detección y solución de valores atípicos es esencial ya que su presencia puede sesgar los resultados de los análisis estadísticos en el conjunto de datos. Esto conduciría a modelos menos efectivos y menos útiles. Es por ello que el objetivo de la detección de valores atípicos es el estudio del impacto de estos valores en la distribución del conjunto de datos a tratar y la consideración del tratamiento de estos, pudiendo conservarlos o eliminarlos para así optimizar el resultado obtenido con los posteriores modelos de aprendizaje automático⁶⁰.

Existen diferentes técnicas para identificar los valores atípicos^{61,62}:

- Detección mediante la desviación estándar: Si los datos siguen una distribución normal, se puede usar la desviación estándar de la muestra para detectar valores atípicos. Se suele determinar un valor atípico cuando supera las tres desviaciones estándar con respecto a la media (en los dos límites).
- Detección mediante Z-score: Para distribuciones sesgadas, se puede usar el valor de Z-score definido como: $z = (x - \mu)/\sigma$, siendo x el valor del dato, μ la media de la distribución y σ el valor de la desviación estándar. En esta técnica se usan tres valores hacia arriba y tres valores hacia abajo como límite de la determinación de valores atípicos.
- Detección mediante rango intercuartil: Se define el rango intercuartil como la cantidad que mide la diferencia entre el primer y el tercer cuartil en un conjunto de datos determinado. Se suele usar el rango entre valores menores al primer cuartil y mayores al tercer cuartil para detectar los valores atípicos. Para este método se suele representar las variables mediante un gráfico de tipo caja y bigotes ya que se muestran directamente los puntos que se salen de los cuartiles anteriormente mencionados.
- Detección mediante percentiles: Parecido a la técnica basada en intercuartiles, se estudia la muestra con respecto a los percentiles y se suele establecer un rango de 0.5 a 99.5. Esta técnica permite estudiar valores atípicos en distribuciones más amplias evitando descartar puntos que no son valores atípicos.
- Uso de algoritmos: Existen diferentes aproximaciones para la detección de valores atípicos mediante el uso de algoritmos. Por una parte, se usan algoritmos de agrupación en grupos que detectan valores atípicos comparando la distancia de los puntos con respecto al centro de los grupos más cercanos. Por otro lado, se pueden usar algoritmos dentro de la familia de árboles de decisión los cuales aíslan las anomalías mediante la asignación de una puntuación a cada valor. Detecta los valores

anómalos ya que suelen tener valores diferentes a los otros puntos.

- Detección mediante proporción de clases: Para el estudio de valores atípicos en variables categóricas, se estudia la frecuencia de cada categoría determinando las clases con menos incidencia como valores atípicos.

Con las técnicas de identificación descritas, se puede proceder al tratamiento de los valores atípicos existiendo tres enfoques diferentes⁶³:

- Recorte o eliminación: Detectando los valores atípicos, se excluyen estos mismos valores de nuestro análisis.
- Limitación: Con esta técnica se establece un límite para definir qué valores se consideran valores atípicos y cuales no.
- Discretización: Se hacen grupos donde se incluyen los valores atípicos en un grupo particular y los obligamos a comportarse de la misma manera que los otros puntos de ese grupo.

Reducción de dimensionalidad

La ejecución y rendimiento de los algoritmos de aprendizaje automático puede verse perjudicados con demasiadas variables de entrada. En una interpretación geométrica de los datos, se considera cada variable (representada por una columna del *dataframe*) como una dimensión dentro de un espacio de n -dimensiones y las observaciones (filas en el *dataframe*) como puntos dentro de esos espacios. Si se tiene un conjunto de datos con muchas variables, se estaría tratando de un espacio de múltiples dimensiones donde resulta difícil distinguir la importancia de las observaciones ya que suponen unos puntos no representativos dentro de ese espacio de n -dimensiones.

Los datos con múltiples variables suelen dar dos tipos de problemas, el primero conocido como *curse of dimensionality* (maldición de la dimensionalidad), donde el modelo pierde rendimiento a medida que van aumentando el número de características. Por otro lado, tenemos que puede ocurrir un sobreajuste del modelo, ajustándose muy bien a los datos de entrenamiento pero no pudiendo generalizar con datos nuevos⁶⁴.

Para mitigar o evitar estos dos problemas, se realiza la reducción de dimensionalidad. Ésta es una técnica utilizada para reducir la cantidad de características en un conjunto de datos mientras se retiene la mayor cantidad posible de información importante. Para la realización de esta técnica, existen dos enfoques posibles⁶⁵:

- Selección de variables: Este método implica la selección de un subconjunto de las variables originales que son más relevantes para el problema en cuestión. Se pretende aplicar la reducción de dimensionalidad, eliminando las variables menos influyentes. Esta técnica se puede utilizar siguiendo tres enfoques diferentes:
 - **Método de filtro**: el cual clasifica las variables según su relevancia.
 - **Método envolvente**: utiliza el rendimiento del modelo como criterio de selección de variables.

- **Método integrado:** combina la selección de características con el proceso de entrenamiento del modelo. Una de las técnicas que se suele usar para la selección de variables es *Random Forest Classifier*.
- Extracción de variables: Este método implica la creación de nuevas variables combinando o transformando las variables originales. Con este método se crea un conjunto de variables que capturan la esencia de los datos originales en un espacio de menor dimensión. Las técnicas que se suelen usar en este caso son el análisis de componente principal (PCA) o análisis de correspondencia múltiple (MCA).

B. Aprendizaje automático

El aprendizaje automático es una rama de la inteligencia artificial y la ciencia de la computación que se enfoca en el uso de los datos y algoritmos con el fin de imitar la forma en que los humanos aprenden, mejorando progresivamente su precisión⁶⁶. Mediante el uso de métodos estadísticos, los algoritmos se entrenan para hacer clasificaciones o predicciones y descubrir información clave en diferentes proyectos de análisis de datos o minería de datos.

Los algoritmos de aprendizaje automático construyen un modelo basado en datos de muestra, conocidos como datos de entrenamiento, para hacer predicciones o tomar decisiones sin estar programados explícitamente para hacerlo. Estos algoritmos se utilizan en una amplia variedad de aplicaciones, como en medicina, reconocimiento de voz, agricultura y visión artificial, entre otros. Dentro del aprendizaje automático se encuentra⁶⁷:

- Aprendizaje automático supervisado: Se define por el uso de conjuntos de datos etiquetados con el fin de entrenar algoritmos para clasificar datos o predecir resultados con precisión. Se conocen los valores correspondientes a la variable respuesta, proporcionando esa información al algoritmo para así aprender la relación entre la variables dependiente y las variables independientes.
- Aprendizaje automático no supervisado: Se utilizan datos sin etiquetar haciendo que la función del algoritmo sea analizar y agrupar conjuntos de datos. Este tipo de aprendizaje automático se usa para el descubrimiento de patrones ocultos o agrupaciones de datos sin necesidad de intervención humana.
- Aprendizaje automático semi-supervisado: Ofrece un término medio entre aprendizaje supervisado y no supervisado. Durante el entrenamiento, utiliza un conjunto de datos etiquetados más pequeño para guiar la clasificación y la extracción de características de un conjunto de datos más grande sin etiquetar. El aprendizaje semisupervisado puede resolver el problema de no tener suficientes datos etiquetados para un algoritmo de aprendizaje supervisado. También ayuda si es demasiado costoso etiquetar suficientes datos.

Debido a las características de cada modelo de aprendizaje automático, cada uno de ellos se aplica para la resolución de diferentes tipos de problemas. Como se ha explicado en el apartado 1.4 de esta misma memoria, el enfoque del aprendizaje automático en el campo de la medicina es la predicción de valores pudiendo así predecir de manera más concreta los resultados de nuevos datos¹¹. Es por ello, que como también se comenta en ese

mismo apartado, en este trabajo se implementan técnicas de aprendizaje supervisado que sirvan para la creación de modelos de predicción. En concreto, se usan los algoritmos de árboles de decisión, bosques aleatorios, máquinas de vectores de soporte y redes neuronales artificiales.

ÁRBOLES DE DECISIÓN:

Los árboles de decisión son un modelo jerárquico que se utiliza en el apoyo de decisiones y que representa las decisiones y sus posibles resultados, incorporando eventos fortuitos, gastos de recursos y utilidad. Este modelo se basa en sentencias de control condicional, son algoritmos normalmente usados en aprendizaje automático supervisado, siendo útil tanto para tareas de clasificación como para tareas de regresión⁶⁸.

Se llaman algoritmos de árboles de decisión debido a la estructura de árbol de la cual se componen, teniendo un nodo raíz, ramas, nodos internos y nodos hoja, formando una estructura jerárquica similar a un árbol. Este algoritmo se basa en la segmentación del conjunto de datos inicial en diferentes conjuntos de datos más simples, dividiendo más en cada paso el conjunto de datos hasta que los resultados son homogéneos o se cumple algún criterio de selección. El criterio para la división de estos datos suele ser la selección de una de las variables con más peso en la predicción creando los nodos hijos con respecto a distintos valores de esa variable formando así un conjunto de nuevas ramas.

Para mejorar la comprensión de la terminología de este algoritmo, se explica la estructura de este mismo:

- Nodo raíz: Corresponde al nodo superior del árbol, siendo este el inicio del algoritmo. Representa el conjunto de datos inicial que añadimos al modelo.
- Nodo decisión/interno: Nodo que simboliza una elección con respecto a una característica de entrada. La ramificación de los nodos internos los conecta a los nodos hoja u otros nodos internos.
- Nodo hoja/terminal: Nodo sin ninguna continuación. Son los nodos donde no es posible una mayor división.
- Sub-árbol/rama: Subsección del árbol de decisión que empieza con un nodo interno y acaba en un nodo hoja.
- Nodo padre: Nodo que se divide en uno o más nodos hijos.
- Nodo hijo: Nodo que emerge cuando el nodo padre se separa.

Árboles de decisión

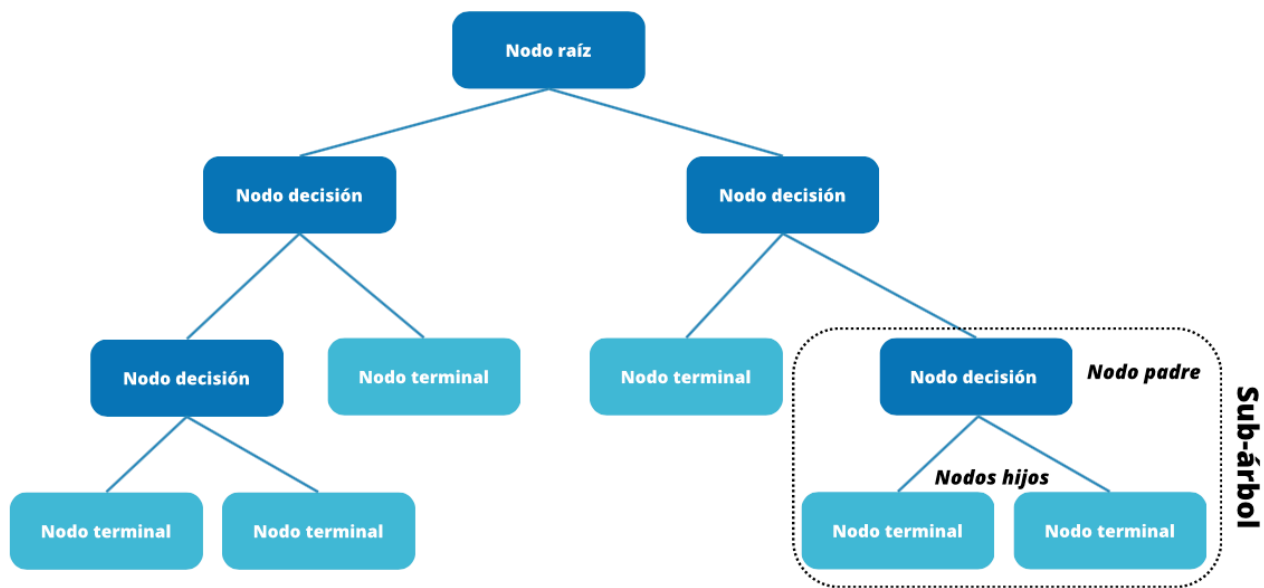


Figura 2.3 Árboles de decisión. Esquema gráfico de la estructura del algoritmo árboles de decisión.

A parte de conocer la terminología de la estructura del propio algoritmo, se procede a explicar términos relacionados con los procesos que se llevan a cabo para la creación de estos árboles mediante la toma de decisiones. En la definición del modelo, se debe tener en cuenta estos parámetros conociendo el conjunto de datos inicial que se tiene ya que son los que marcan cómo se construye el árbol. Estos parámetros son⁶⁹:

- División: Proceso de división de un nodo en dos o más subnodos usando un criterio de división.
- Impureza: Medida de la homogeneidad de las variables en un conjunto de los datos. Se refiere al grado de azar o impureza de una parte del conjunto de datos.
- Varianza: Este atributo mide cómo las variables de predicción y la variable objetivo cambian en diferentes muestras del conjunto de datos. Se usa para problemas de regresión.
- Ganancia de información: Es una medida de la reducción de la impureza conseguida mediante el recorte de un conjunto de datos en una variable en concreto dentro del árbol de decisión. Se usa para determinar qué variable es la más informativa para poder dividir el nodo.
- Poda: Proceso donde se eliminan las ramas del árbol que no aportan información adicional o que pueden condicionar a un sobreajuste.

En el apartado de metodología se ampliará la información sobre los hiperparámetros más específicos en el algoritmo desarrollado en este trabajo.

BOSQUES ALEATORIOS:

Bosques Aleatorios es un algoritmo basado en métodos de agrupación enfocados en agrupar diferentes árboles de decisión. Este algoritmo se aplica en modelos de aprendizaje automático supervisado que se usa ampliamente en problemas de clasificación y regresión.

La estructura de bosques aleatorios se basa en un conjunto de árboles de decisión individuales, entrenando cada uno de ellos con una muestra aleatoria que se extrae del conjunto de datos original mediante la técnica de *bootstrapping*, técnica basada en un muestreo aleatorio con reemplazo⁷⁰. Cada árbol de decisión actúa como un experto en su campo, dando unos resultados de clasificación de los datos ligeramente diferentes. La predicción final se realiza mediante el cálculo de la predicción de cada árbol individual y tomando el valor con más peso.

Debido a que este algoritmo es un conjunto de árboles de decisión, se consideran la misma estructura y los mismos parámetros explicados en el apartado anterior, estando la estructura formada por distintos nodos y ramas y los parámetros definidos por impureza, varianza, etc⁷¹.

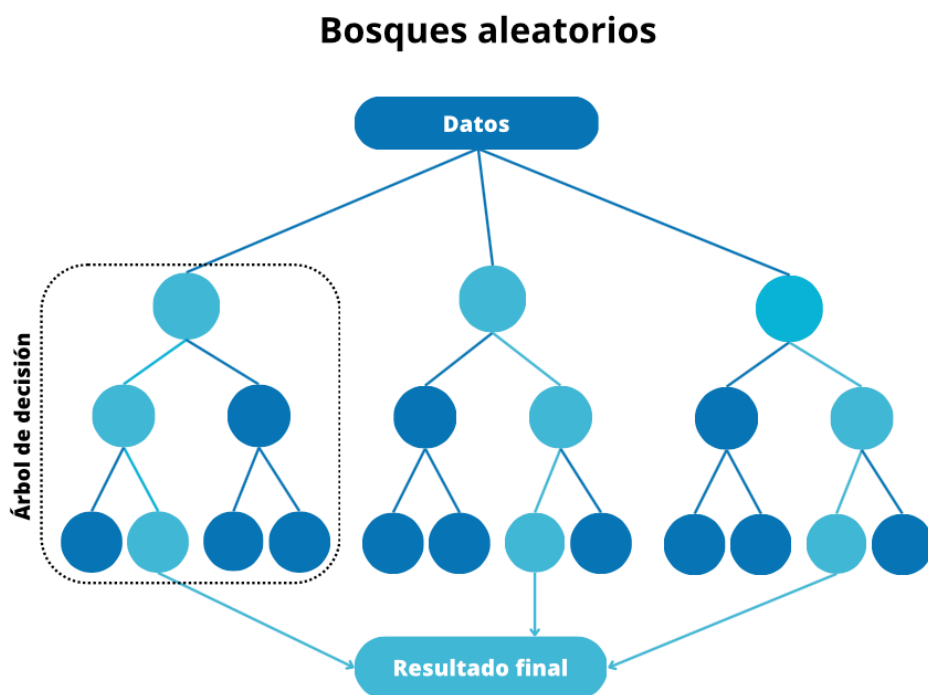


Figura 2.4 Bosques aleatorios. Esquema gráfico de la estructura y funcionamiento del algoritmo tipo Bosques aleatorios.

MÁQUINAS DE VECTORES DE SOPORTE:

Las máquinas de vectores de soporte es uno de los algoritmos más utilizados dentro del mundo del aprendizaje automático. Es poderoso, fácil de explicar y generaliza bien en muchos tipos de datos siendo tanto datos categóricos como numéricos. Suele usarse para problemas de clasificación aunque también puede usarse en problemas de predicción⁷².

Este algoritmo representa cada observación como puntos en un espacio, buscando cuál es el mejor hiperplano para la separación de los puntos en diferentes clases. Este hiperplano se define por el vector que determina la distancia más corta de dos puntos de dos clases diferentes, llamando a este vector, vector soporte. Cuando las nuevas muestras se ponen en correspondencia con dicho modelo, en función de los espacios a los que pertenezcan, se clasifican conforme a una clase u otra. Es por ello que los componentes principales en este algoritmo son: vector de soporte, margen máximo e hiperplano (véase Figura 2.5).

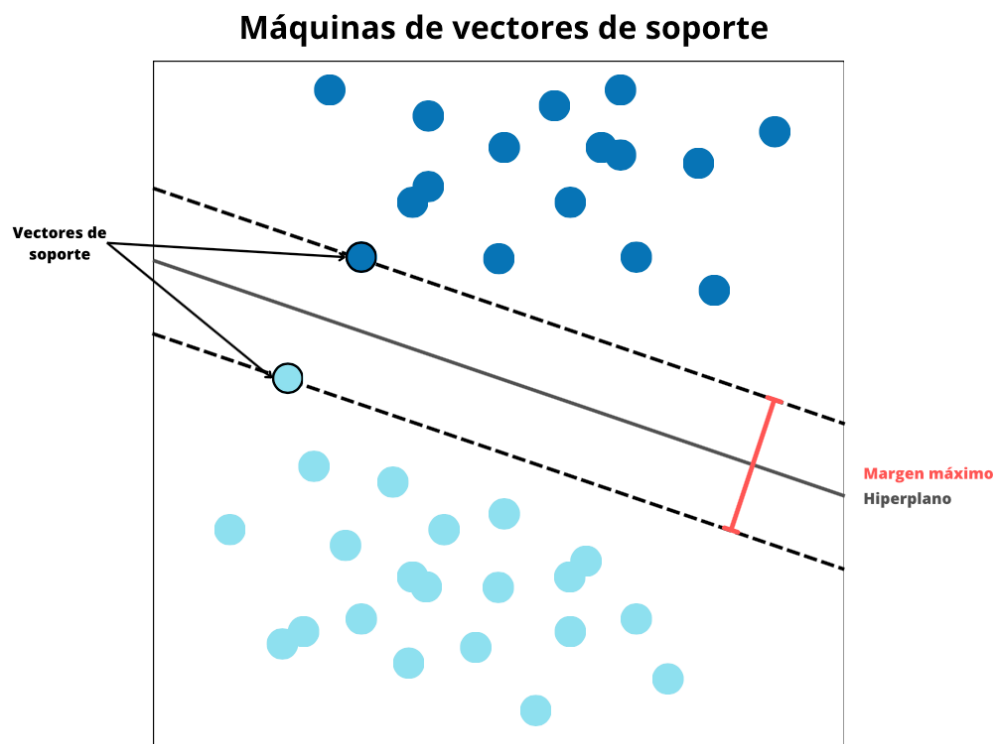


Figura 2.5 Máquinas de vector de soporte. Esquema gráfico de definición de máquinas de vectores de soporte y elementos principales: vectores de soporte, margen máximo e hiperplano.

Además de los conceptos a nivel estructural del algoritmo, también se deben conocer las características que definen un modelo de máquinas de vectores de soporte⁷³.

- **Kernel:** Es una función que permite proyectar los datos originales desde su espacio original a un nuevo espacio transformado, pudiendo así facilitar una separación lineal de los datos. Se suelen utilizar kernel lineal, radial o polinomial.
- **Regularización:** Es el valor de la distancia dentro del margen máximo que se permite sobrepasar sin contar como una penalización. Se mide por el valor del parámetro C . Para valores mayores de C , se aceptará un margen más pequeño si la función de decisión clasifica mejor todos los puntos de entrenamiento correctamente. Una C más baja fomentará un margen más grande, por lo tanto, una función de decisión más simple, a costa de la precisión del entrenamiento.

- Gamma: Define la influencia de un ejemplo de entrenamiento, donde los valores bajos significan un ajuste más libre a los datos de entrenamiento y valores altos, un ajuste excesivo al conjunto de datos de entrenamiento. De hecho, si gamma es un valor muy grande, el radio del área de influencia de los vectores de soporte solo incluye el vector de soporte en sí mismo y ninguna cantidad de regulación con C puede evitar el sobreajuste.

REDES NEURONALES ARTIFICIALES:

Las redes neuronales artificiales son un método dentro de la inteligencia artificial que enseñan al ordenador a procesar datos de una forma inspirada en el funcionamiento del cerebro humano. La idea es replicar el comportamiento de las millones de neuronas interconectadas que existen en el cerebro.

En este caso, cada neurona está representada por un nodo pudiendo diferentes nodos agruparse formando una matriz, denominada capa. Una estructura de una red neuronal simple, está compuesta por⁷⁴:

- Capa de entrada: El conjunto de datos a procesar será introducido a la red neuronal mediante la capa de entrada. Esta capa procesa, analiza y categoriza este conjunto de datos, pasándola después de su análisis a la siguiente capa.
- Capa oculta: Estas capas obtienen información de las capas anterior, siendo o bien la capa de entrada u otras capas ocultas. Cada capa oculta analiza el resultado de la capa anterior, lo procesa y lo vuelve a enviar a la siguiente capa.
- Capa de salida: Esta capa muestra el resultado final del procesamiento realizado por la red neuronal. Puede estar compuesta por uno o más nodos dependiendo de si se ha analizado un problema de clasificación binaria (dando un nodo de salida) o problemas de clasificación múltiples (teniendo más de un nodo de salida).

Dependiendo de la forma en que la información fluya dentro del modelo de redes neuronales, se pueden diferenciar distintos tipos de redes neuronales⁷⁵:

- Red neuronal del tipo hacia delante (*feedforward*): El procesamiento de los datos se da en una sola dirección, desde la capa de entrada hasta la capa de salida pasando por todas las capas ocultas. Cada nodo se conecta con todos los nodos de la capa siguiente.
- Red neuronal de propagación hacia atrás (*backpropagation*): Este tipo de algoritmo permite que la información de la red vuelva hacia atrás para poder calcular el gradiente de error que se va acumulando.
- Red neuronal tipo convolucional: Las neuronas funcionan como campos perceptivos parecidos a las neuronas de la corteza visual del cerebro biológico. Se añaden capas de convolución que permiten la extracción de características de los datos iniciales. Luego se añaden las capas de reducción que permiten ir simplificando los datos iniciales para poder reducir la dimensionalidad de los datos iniciales. Se suelen usar en visión artificial.

El funcionamiento de la red neuronal viene marcada por la definición de funciones que permiten controlar tanto la propagación de la información dentro de la red como la predicción de la red con respecto al valor final. En este algoritmo, se controla la importancia de cada información que recibe cada neurona de cada capa. Para esclarecer estos conceptos, se puede observar el diagrama descrito en la Figura 2.6 mostrando la “anatomía” de una neurona dentro de una red neuronal.

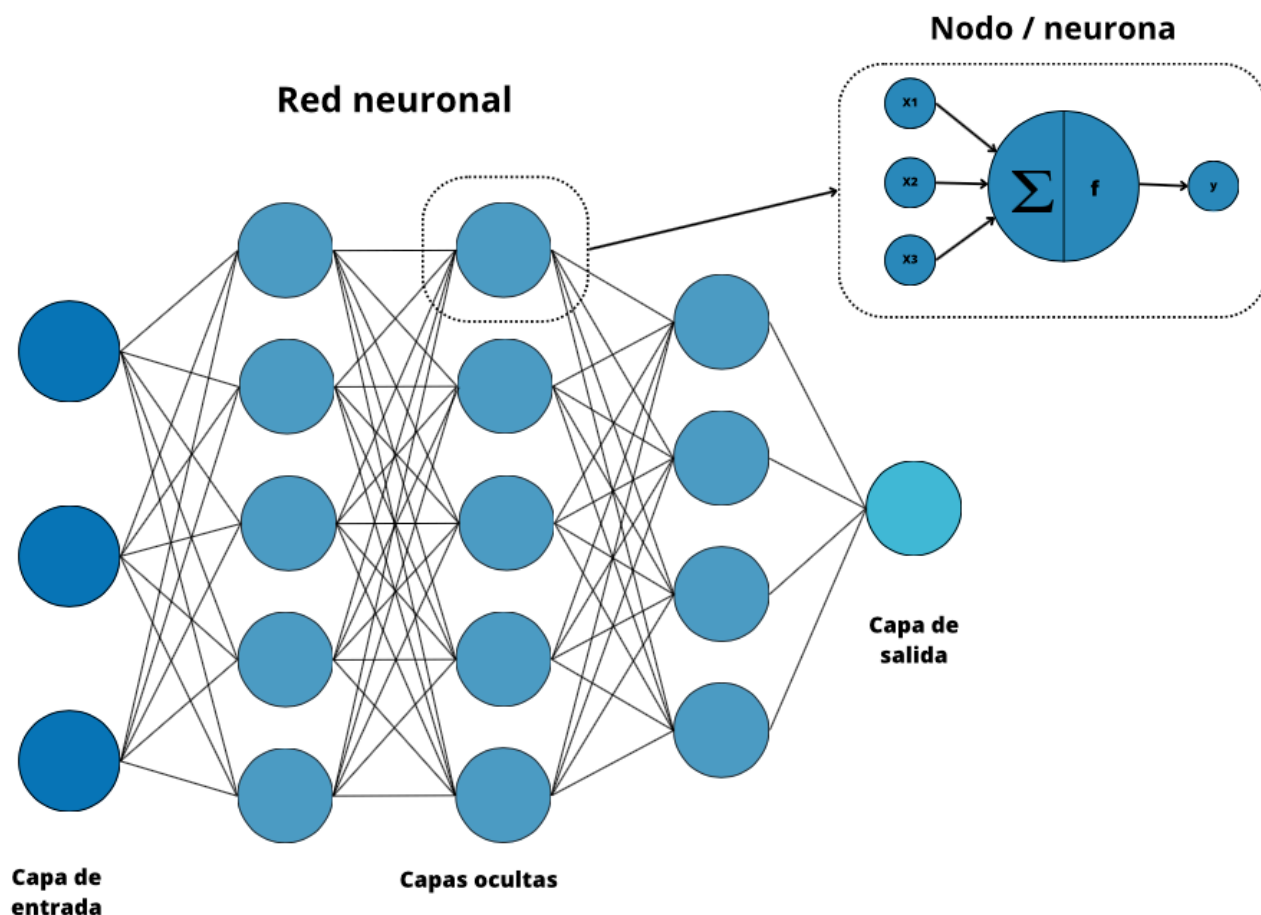


Figura 2.6 Red neuronal artificial. Esquema gráfico de la arquitectura de red neuronal artificial y la estructura interna de un nodo.

Como se observa, la neurona capta las señales de entrada (variables x) de cada una de las neuronas anteriores como si se tratase de las dendritas de las neuronas reales. Cada señal de entrada es ponderada mediante una variable, llamada peso (variable w) de acuerdo con la importancia que se le da a esta entrada. Todas las señales de entrada son sumadas en el “cuerpo celular” y la señal resultante se transmite a la siguiente neurona de acuerdo con la función de activación escogida⁶⁷.

En la definición del modelo a nivel de computación, se deben definir las características tanto a nivel neuronal como a nivel de estructura de red. Estos parámetros a tener en cuenta son^{76,77,78}:

- Función de activación: Función que transforma todas las señales de entrada de una neurona y las convierte en una señal de salida. Pueden ser:

- **Función activación Ridge:** Funciones multivariantes que se aplican en combinaciones lineales de las variables de entradas. Las más usadas son funciones ReLU, lineal o logística/sigmoidal.
 - **Función activación radial:** Funciones de tipo radial para modelos con variables de entrada no lineales. Las funciones Gauss o multicuadrática.
 - **Función activación “folding”:** Se usan en capas de agrupación en el caso de las redes neuronales convolucionales. La función más conocida es la función de activación *softmax*.
- Peso: Coeficientes que muestran la importancia de la información de cada entrada a una neurona concreta. La determinación de estos pesos viene dada por diferentes algoritmos como son el gradiente de descenso o ajuste de pesos por “fuerza bruta”, entre otros.
- Función de coste: Función encargada de cuantificar la distancia entre el valor real y el valor predicho por la red, siendo un mecanismo de determinación del error de predicción. Existen diferentes tipos de función de coste:
- **Raíz cuadrada media (RMSE):** Medida de precisión calculada como la raíz cuadrada media de los residuos. Se suele usar para la optimización de regresiones en general.
 - **Error absoluto medio (MAE):** Medida de precisión calculada como la suma media de los valores absolutos de los errores.
 - **Entropía cruzada categórica:** Medida de precisión para variables categóricas.
 - **Entropía cruzada binaria:** Medida de precisión para variables categóricas de carácter binario.

EVALUACIÓN DE LOS ALGORITMOS:

El último paso que se realiza en todos los algoritmos de aprendizaje automático es la evaluación de los modelos. Existen diferentes opciones dependiendo del tipo de datos con el que se trabaje^{67,79}.

Precisión de clasificación

Corresponde al ratio de las predicciones realizadas correctamente con respecto al número de predicciones totales. Es el método más usado para la evaluación de los modelos en general; no obstante, no se recomienda su uso en conjuntos de datos no balanceados.

$$Precision = \frac{Numerodeprediccionescorrectas}{Totaldepredicciones}$$

Matriz de confusión

Matriz que muestra los valores de falso positivo, falso negativo, positivo verdadero y negativo verdadero. La precisión del modelo se puede obtener calculando la media de los valores de verdaderos positivos y verdaderos negativos con respecto a las muestras totales.

	Predicción: Sí	Predicción: No
Valor real: Sí	Positivo verdadero (<i>True positive, TP</i>)	Falso negativo (<i>False Negative, FN</i>)
Valor falso: No	Falso positivo (<i>False Positive, FP</i>)	Negativo verdadero (<i>False Negative, FN</i>)

La precisión del modelo se puede obtener calculando la media de los valores de verdaderos positivos y verdaderos negativos con respecto a las muestras totales.

$$Precision = \frac{TruePositive + TrueNegative}{Total}$$

Área bajo la curva (AUC) - Curva ROC

Son medidas usadas para problemas de clasificación binarios. La curva ROC es una curva de probabilidad donde se representan los valores del ratio de verdaderos positivos con respecto al ratio de falsos positivos. AUC es el valor debajo de la curva ROC e indica la capacidad de distinción entre las clases. Cuanto mayor sea el valor de AUC, mejor será el modelo para predecir cada clase.

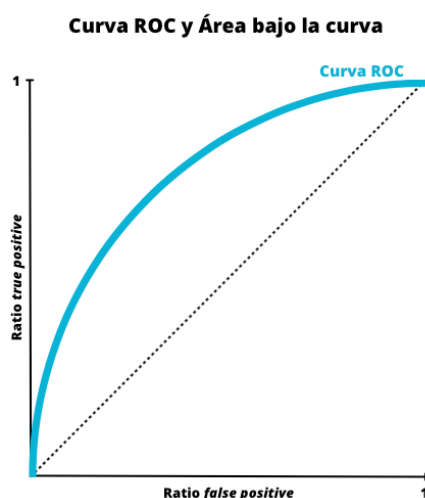


Figura 2.7. Curva ROC y Área bajo la curva. Representación gráfica de los conceptos curva ROC y valor AUC.

Valor f1

Se considera la medida media entre la precisión y la exactitud. Esta medida muestra la precisión y robustez del modelo. Siendo su fórmula matemática:

$$F1 = 2 * \frac{1}{\frac{1}{precision} + \frac{1}{robustez}}$$

Se define precisión como el número de verdaderos positivos dividido entre el número de los resultados positivos generales (verdaderos positivos más falsos positivos). Y robustez como el número de verdaderos positivos entre el número de las más muestras identificadas como positivas (verdaderos positivos y falsos negativos).

3. MATERIALES Y MÉTODOS

3.1. ANÁLISIS DE DATOS

A. Análisis exploratorio de datos

En este paso el objetivo principal es la colección, limpieza y transformación de nuestro conjunto de datos. Este procedimiento se inicia con un **análisis exploratorio** del conjunto de datos MSF. En este apartado, se carga el conjunto de datos y se comprueba que se haya importado de manera correcta usando funciones que vienen dadas por paquetes básicos de Python como es Pandas⁸⁰, paquete específico para el manejo de *dataframes*. También se comprueba qué tipo de variables componen el conjunto de datos, siendo este un factor clave para la determinación de los siguientes pasos.

Antes de pasar al siguiente apartado de **preprocesamiento** de estos datos, se procede a la obtención de la variable dependiente (Y) que describe si se tiene un parto prematuro o no (corresponde a los valores de la variable “*PreTerm*” del conjunto de datos original). El resto de variables, se definen como la matriz de variables independientes X. Para ello, se siguen usando funciones que vienen dadas por librerías comunes de Python, como la anteriormente mencionada Pandas, Matplotlib⁸¹ para la realización de gráficos o Numpy⁸². Se comprueba la distribución de la variable respuesta viendo si las proporciones de cada nivel de esta variable son equivalentes o no se encuentran balanceadas usando gráficos y cálculos de proporción. Este paso es importante ya que se debe tener en cuenta en la predicción y análisis de resultados, evitando obtener conclusiones que no son correctas.

Se separan los datos en entrenamiento y test usando la función *tran_test_split* del paquete *Sklearn*⁸³. El paso de la creación de entrenamiento y test es esencial dentro de la práctica de aprendizaje automático, ya que los algoritmos necesitan realizar una evaluación del modelo obtenido. Una vez que se ha procesado un modelo mediante el uso del conjunto de entrenamiento, se realiza una prueba de este haciendo predicciones con el conjunto de prueba. Debido a que los datos en el conjunto de prueba ya contienen valores conocidos para el atributo que desea predecir, es fácil determinar si las predicciones del modelo son correctas⁶¹.

B. Preprocesamiento de datos: Estudio de valores nulos

Teniendo ya este conjunto de datos preparado para su manejo, se inicia el paso de preprocesamiento de datos. En primer lugar, se estudia la existencia de valores nulos y el tratamiento de estos.

Este paso es uno de los básicos en el análisis de datos, debido a que la mayoría de algoritmos de aprendizaje automático no funcionan si el conjunto de datos tiene valores nulos. Además si se trabaja con valores nulos se puede terminar creando un modelo sesgado que nos genere resultados incorrectos o con falta de precisión estadística.

Conociendo el tipo de valores nulos que existen, se procede a la identificación y manejo de estos teniendo en cuenta que cada tipo de valor nulo requiere de un tratamiento específico. Para la identificación de estos valores se usan funciones dadas por el paquete Pandas de Python como por ejemplo `.isnull()`. Con ello, se comprueba la existencia de valores nulos, la cantidad y en qué variables se encuentran estos valores. En el caso de este conjunto de datos, identificamos los valores nulos del tipo MAR siendo valores faltantes para algunas de las variables.

Conociendo el tipo de valores nulos, la cantidad y las variables que los contienen, se procede a la elección del método de mitigación de estos valores nulos. En este trabajo la eliminación de las variables u observaciones con valores nulos no se aplica debido a que supone la pérdida de más del 50% de observaciones y por la imposibilidad de la eliminación de ciertas variables esenciales en la definición del modelo. Es por ello, que se decide aplicar técnicas de imputación para sustituir estas variables por algún valor que permita la realización de los modelos de la manera más correcta posible.

En el enfoque de este trabajo se han aplicado diferentes técnicas de reemplazo de valores nulos dependiendo del tipo de variables. Para variables numéricas se ha optado por el valor de la media (variable numérica continua) o el valor común (variable numérica discontinua). Por otra parte, para las variables consideradas categóricas se han aplicado tres estrategias: la creación de una nueva clase (se determina el valor 10 como clase nueva), valor común o uso del algoritmo kNN para la imputación de estos valores⁸⁵.

C. Preprocesamiento de datos: Estudio de valores atípicos

En este trabajo se opta por el uso de la técnica de limitación ya que se va a considerar la opción de mantener los valores atípicos o no; conociendo los valores que sobrepasan un umbral se puede decidir si estos supondrían un problema en el modelo⁸⁶.

Por otro lado, se opta por la opción de uso de algoritmos (en concreto PCA y *Isolation Tree*) para la detección de valores anómalos. Se usará la librería específica de PyOD⁸⁷ que está diseñada para la aplicación de diversos algoritmos específicamente enfocados en la detección de valores atípicos. Se descartan las otras opciones debido a la dificultad de estudio de cada uno de los gráficos o valores estadísticos para las 123 variables que se manejan en este trabajo.

Cabe destacar, que aún sabiendo que PCA es un método de reducción de dimensionalidad, en la librería PyOD aparece como opción aplicada para la detección de valores atípicos. Es así que las características de este modelo aparecerán adaptadas a la funcionalidad de este método. Se usa el error de reconstrucción producido al revertir la

reducción de dimensionalidad como medida de detección de valores atípicos. Las observaciones más próximas al promedio son las que mejor están proyectadas y las mejores reconstruidas, teniendo un valor más pequeño de error de reconstrucción. En cambio, los valores atípicos serán aquellos que tengan un error de reconstrucción más grande, siendo las observaciones más alejadas del promedio.

Por el contrario, *Isolation Forest* sí que es un algoritmo que está diseñado para la detección de valores atípicos. La base teórica es el uso de modelos que intentan aislar las anomalías del resto mediante el uso de un conjunto de árboles de decisión. Entonces, se selecciona una característica y se hace una división aleatoria de los datos entre valores mínimo y máximo.

En la aplicación de los dos modelos se siguen tres pasos:

- Desarrollo del modelo: Se cargan las librerías de cada uno de los métodos (PCA o *Isolation Forest* de la librería PyOD), se define el modelo mediante las funciones específicas del paquete y se calculan los niveles de predicción y las puntuaciones de los valores atípicos. Además, se muestran los parámetros de cada uno de los modelos para la comprobación de este.
- Cálculo de treshold: Para conocer el valor umbral que determine qué puntos son considerados valores atípicos o no, se usa la función *threshold* de los dos métodos implementados. Para poder realizar la comprobación visual del umbral de cada uno de los modelos, se muestra un histograma con los valores de outliers.
- Determinación de grupos normal y outlier: Con el fin de conocer qué valores son considerados atípicos según el umbral establecido en el paso anterior, se dibuja una tabla con el número de observaciones consideradas normales en comparación con valores atípicos. Además, se añade el porcentaje del total de *outliers* y la puntuación de anomalía para cada variable pudiendo comprobar la diferencia existente entre las observaciones normales y las atípicas.

Por último, se calcula una matriz de confusión para la comprobación de la predicción de los valores atípicos pudiendo así valorar si los modelos funcionan correctamente o no.

D. Preprocesamiento de datos: Reducción dimensionalidad

En este trabajo se opta por el uso de técnicas del tipo selección de variables. El objetivo es obtener una selección de las variables más significativas para su posterior uso en la creación del cuestionario. En concreto, se aplicará el algoritmo *Random Forest Classifier* ya que su implementación en Python se puede realizar de manera directa y sencilla mediante la librería *sklearn.ensemble*.

Cabe destacar que este algoritmo no está diseñado específicamente para la reducción de dimensionalidad. No obstante, sí permite obtener la importancia relativa de cada variable en el modelo pudiendo así seleccionar las variables más relevantes.

Los pasos a implementar en este procedimiento consisten en:

- Preparación del conjunto de datos: Se elimina la variable etiqueta (*Mother_UID*) para evitar problemas de ejecución del algoritmo.
- Definición del modelo y obtención de las variables: Se usa la función *RandomForestClassifier* para la definición del modelo en los conjuntos de datos pertinentes. Seguidamente se crea una variable con el nombre de las variables de los modelos usando la opción *.columns.values* y se adjudican estos valores en un *dataframe*. En este mismo *dataframe* se añaden los valores de importancia del modelo de cada variable. Para mejorar la visualización, estos valores se ordenan de mayor a menor para obtener los diez valores más importantes como los primeros valores del conjunto de datos.
- Comprobación de la importancia de las variables: Se describe una gráfica de tipo histograma con la librería *seaborn* que muestra la importancia de cada variable.
- Definición del conjunto de datos reducido: Conociendo las diez variables con más relevancia dentro de los modelos, seleccionamos estas en el conjunto de datos *X_train1*, *X_train2* y *X_train3*. Así obtenemos los conjuntos de datos reducidos con solamente 10 variables.

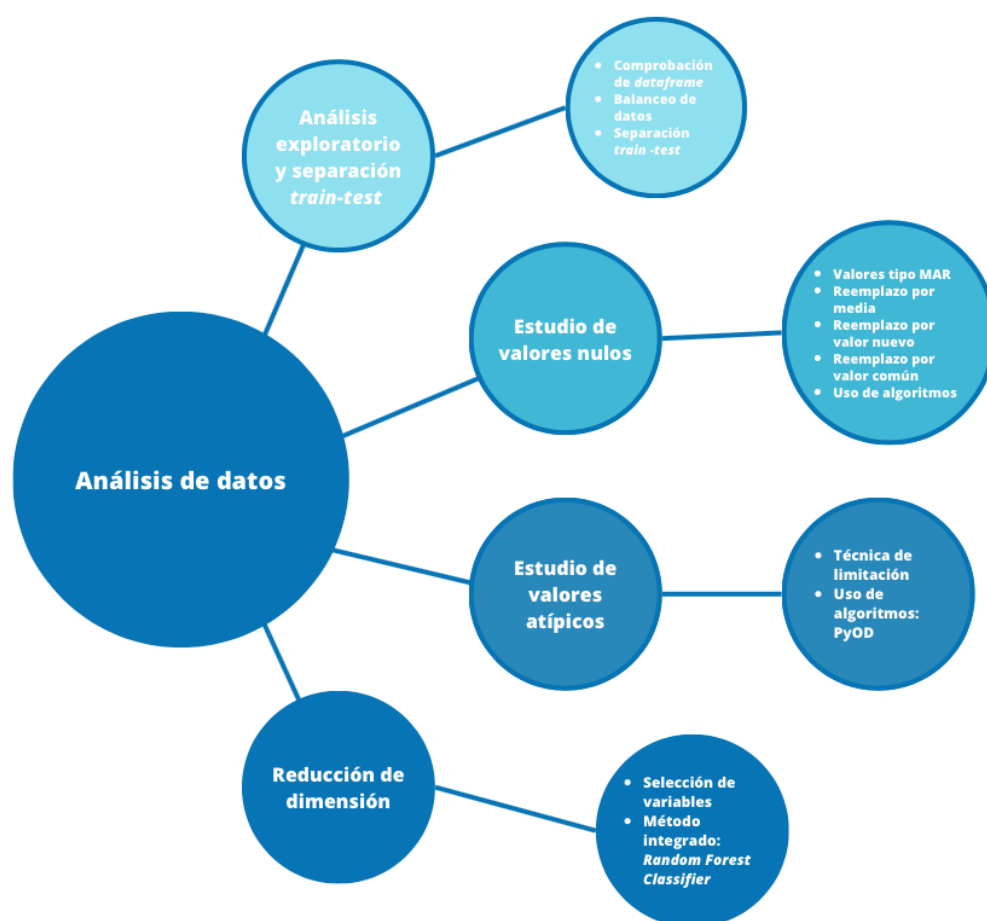


Figura 3.1 Análisis de datos. Mapa conceptual de pasos seguidos en el análisis de datos y elecciones de metodología en cada paso.

3.2. APRENDIZAJE AUTOMÁTICO

A. Árboles de decisión

En este caso, se aplica este algoritmo para poder resolver el problema de predicción planteado en la hipótesis de trabajo. Para la implementación del código, se usa la librería de *Sklearn* con los paquetes específicos de este algoritmo (*DecisionTreeClassifier*, *metrics*, *tree*). La opción inicial será el uso de los hiperparámetros predefinidos; no obstante, se procede a explicar los parámetros más importantes que se modifican en caso de requerir el paso de optimización del modelo:

- Criterio: Este parámetro permite la elección de diferentes criterios de selección de las variables. Se puede elegir entre:
 - *Índice Gini*: Evalúa la precisión de la división entre los grupos clasificados. Tiene un rango entre 0 y 1, donde 0 corresponde a las observaciones de una sola clase y 1 a una distribución aleatoria de los elementos dentro de las clases. El objetivo es obtener un índice de Gini lo más bajo posible. Es el parámetro que viene dado por defecto.
 - *Entropía*: Es la medida del grado de aleatoriedad o incertidumbre en el conjunto de datos. Mide la aleatoriedad en función de la distribución de las etiquetas de clase en el conjunto de datos.
- Splitter: Este parámetro nos permite elegir la estrategia de división. Las estrategias admitidas son “mejor” para elegir la mejor división y “aleatoria” para elegir una división aleatoria de los datos.
- Max_depth: Profundidad máxima del árbol. Si es ninguno, los nodos se expanden hasta que todas las hojas contienen menos de *min_samples_split*.

Los pasos seguidos para este modelo son la importación de las librerías anteriormente mencionadas, seguidamente se define el modelo mediante la función *DecisionTreeClassifier*. Se entrena el modelo para cada conjunto de datos usando la función *fit*, luego se calcula la predicción para los datos test usando la función *predict*. Por último, se calculan los valores de f1 y AUC mediante las funciones *f1_score* y *roc_auc_score* de la librería *sklearn.metrics*.

Para obtener una imagen del modelo, se muestra un gráfico del árbol mediante la librería *graphviz*.

B. Bosques aleatorios

Para este algoritmo, también se usa la librería de *Sklearn* de Python, en concreto el paquete *RandomForestClassifier* para la definición del modelo de Bosques aleatorios. No obstante, en la definición de este modelo mediante *sklearn* se presta como requisito la transformación de las variables categóricas a variables tipo *dummy*. Es por ello, que se añade un paso adicional usando la función *get_dummies* del paquete *Pandas*.

Es cierto, que los modelos de bosques aleatorios no necesitan explícitamente de la transformación de estos valores categóricos como variables *dummy*, ya que como se ha demostrado en varios artículos^{88,89} no hay una diferencia clara entre la capacidad de predicción para modelos con las variables categorizadas o no. No obstante, en la librería *scikit-learn* es un requisito y por ello se realiza este paso.

Por otra parte, se procede a explicar los hiperparámetros de este algoritmo debido a su posible modificación en el caso de añadir un paso de validación del modelo. Estos hiperparámetros son:

- *N_estimators*: Número árboles construidos antes del promedio de las predicciones. El incremento de este parámetro mejora la predicción del modelo, pero aumenta el coste computacional.
- *Max_features*: Número máximo de variables del modelo que se consideran para dividir un nodo.
- *Mini_sample_leaf*: Número mínimo de hojas que se requieren para dividir un nodo interno.
- *Criterion*: Parámetro establecido para la selección de diferentes criterios de selección de las variables. Se puede escoger entre Índice de Gini o Entropía.
- *Max_leaf_nodes*: Número máximo de nodos hoja en cada árbol.

Los pasos para la definición de este modelo son los mismos seguidos en el modelo de árboles de decisión. Se usan las funciones *fit* para la definición del modelo y *predict* para la predicción respecto al modelo del paquete *RandomForestClassifier*. Se calculan los valores f1 y AUC con las mismas funciones de *Sklearn*. Por último, se muestra un gráfico de cada modelo usando la librería *graphviz*.

C. Máquinas de vectores de soporte

Con respecto a la aplicación de este modelo a los datos de este trabajo, se usa la librería *sklearn* de Python, importando el paquete específico de *svm*.

Para realizar una correcta aplicación de este algoritmo, se procede al escalado de los datos usando el paquete *MinMaxScaler* de la librería *sklearn.preprocessing*. Se decide seguir este método de normalización debido a no poder asegurar que todas las variables siguen una distribución normal, imposibilitando esto el correcto funcionamiento del escalado usando el paquete *StandardScaler*.

Por otro lado, con el fin de reducir el tiempo y coste de computación, se decide aplicar la función *GridSearchCV* de la librería *skelearn.model_selection* para la búsqueda de los mejores valores de cada hiperparámetro dando así varios valores de C (0.1, 1, 10, 100), gamma (1, 0.1, 0.01, 0.001) y kernel (lineal o rbf). Mediante la función *.fit* se adapta el modelo a estos hiperparámetros combinándolos de forma aleatoria para conocer cuáles son los que mejor funcionan. Para evaluar los resultados de esta búsqueda de hiperparámetros, se muestran los resultados mediante las funciones *.best_params* y *.best_estimators*. Además, se calculan los datos de clasificación de predicción del modelo con mejor ajuste, pudiendo observar el valor de precisión, *recall*, f1 y *support* para cada nivel de la variable Y.

Conociendo los hiperparámetros más adecuados para cada modelo, se define el modelo usando la función *svm* y especificando los hiperparámetros de este. Luego se entrena el

modelo para cada conjunto de datos y se calcula la predicción del modelo con respecto a los valores de test.

Para evaluar cada modelo, se calcula el valor de f1 y AUC usando la función de *Sklearn*.

D. Redes neuronales artificiales

Para la realización de este trabajo se ha escogido el modelo de redes neuronales artificiales debido a la capacidad de clasificación y predicción que posee este algoritmo. Además, se valora la adaptabilidad de este a diferentes datos como datos categóricos o datos numéricos.

Para la aplicación de este algoritmo, se ha usado la librería *tensorflow* y *keras* de Python. Antes de empezar la definición del modelo de red neuronal, también se escalan las variables del conjunto de datos y se convierte el conjunto de datos a matriz tipo *numpy*. En este caso, se usan los conjuntos de datos adaptados para el algoritmo anterior.

Para la definición de la red neuronal, se opta por una red de tipo hacia delante (*feedforward*) debido a la simplicidad y longitud del conjunto de datos a aplicar. No se considera el uso de modelos más complejos como el tipo hacia atrás (*backforward*) ya que no se requiere de la comprobación de los errores acumulados de la red durante el procesamiento de información debido a que no se desarrolla una red con un alto número de capas.

El primer paso que se sigue a nivel de código es la definición del modelo, estableciendo las medidas que se quieren obtener: precisión y *recall* para el cálculo de f1 y auc. Por otro lado, se especifica la arquitectura de la red neuronal junto con la función *.compile* y sus características.

Siendo así, se describe una red neuronal con una capa de entrada, tres capas ocultas y una capa de salida. La densidad de cada capa neuronal se decide mediante la lectura de bibliografía correspondiente, sabiendo que no existe un método correcto para la determinación del número de neuronas^{67,90}.

Por otro lado, las funciones de activación de las capas ocultas se definen de tipo ReLU, esta función no activa todas las neuronas a la vez rebajando así el coste computacional de la red neuronal. Para la capa de salida, la función de activación se ha seleccionado de tipo sigmoide debido a que esta función asume que las probabilidades de salida deben relacionarse sólo con una clase, dando así una precisión más exacta de la variable dependiente. La función de coste se asume de tipo entropía cruzada binaria ya que la variable independiente solamente tiene dos valores posibles (0-1). Además, se añade una capa de tipo *dropout* y una variable para definir el *bias* de cada red neuronal.

El siguiente paso es la construcción del modelo. Para ello, se definen los valores básicos del modelo siendo *batch* y *epoch*. Se usa un valor de batch de 2048 para garantizar que cada valor de la variable tipo de parto tenga una probabilidad decente de contener algunas muestras de tipo parto prematuro. Si el tamaño del estudio fuera demasiado pequeño, probablemente no se tendría los valores de partos prematuros para poder aprender. Se añade la opción de *early stopping* para cuando el modelo llegue a un valor constante de f1 o pérdida de error, se pare de calcular el modelo, evitando así problemas como *overfitting*.

Debido a que partimos de un conjunto de datos que no está equilibrado, se añade un paso de corrección de *bias* intentando mejorar el ajuste del modelo al conjunto de datos. Para ello, se usa la función de $bias = \log(\text{positivo}/\text{negativo})$, siendo positivo y negativo los valores de las clases de la variable dependiente Y. Para la comprobación del valor del bias correcto, se muestran los valores de *loss* anterior y posterior a la corrección de bias, junto con unos gráficos comparativos de estos valores.

Por otro lado, al tener una variable dependiente desequilibrada definiremos los pesos de cada clase para poder igualar la importancia de cada nivel de la variable Y. Para ello, se calculan los pesos para cada uno de los niveles (calculando la distribución de estos con respecto al total). Con los pesos definidos, se vuelve a calcular el modelo añadiendo los pesos con el hiperparámetro *class_weight*. Cabe destacar, que en estos modelos también se añade un conjunto de datos de validación (hiperparámetro *validation_split*) para poder optimizar la red neuronal.

Por último, se calculan los valores de AUC y f1 para cada uno de los modelos. En el caso de las redes neuronales, el cálculo de f1 no está determinado por ninguna función. Es por ello que se calcula de manera manual mediante los valores de precisión y *recall*.

E. Evaluación de algoritmos

En el presente trabajo, se decide usar para todos los algoritmos los valores de **AUC** y el **valor f1** para la evaluación de los modelos. La obtención de estos valores permitirán la comparación y decisión de qué modelo es el más adecuado para el conjunto de datos. El valor de precisión se descarta debido a no tener datos balanceados y la matriz de confusión se deja solamente para mediciones concretas dentro de pasos concretas de alguno de los modelos.

3.3. PRODUCTOS

A. Aplicación web:

El desarrollo web se define como un procedimiento de creación y mantenimiento de un sitio web que resulte funcional en internet a través de diferentes lenguajes de programación como html o Javascript⁹². A parte de los lenguajes de programación más clásicos, también existen *microframeworks* que permiten el desarrollo web de manera más sencilla. En el caso de Python se ha desarrollado Flask, un *microframework* que permite la creación de aplicaciones web pudiendo trabajar con bases de datos definidas en Python, facilitando la creación de una aplicación web solamente usando un documento de este mismo lenguaje⁹³. También, ofrece la opción de *Bootstrap* para el desarrollo de la interfaz gráfica de la aplicación web.

Para el desarrollo de la aplicación web, se decide la implementación de este método ya que permite la conexión de los datos manipulados durante el anterior procedimiento, permitiendo cargar el modelo final de aprendizaje automático a la aplicación web. Además, se evita el uso de lenguajes adicionales como MySQL para la manipulación de estos datos ya que Flask permite vincular la vista de html directamente con el modelo de datos.

Para la importación del modelo de aprendizaje automático, se usa la librería *pickle* que permite a cualquier objeto ser serializado y almacenarlo hasta la deserialización del objeto en otro archivo Python⁹⁴. El estilo y las preguntas del cuestionario se plasmarán en la web mediante el documento de html.

Para la construcción de la página web se usa el lenguaje html pudiendo definir el cuestionario que se pasa a las/los usuarias/usuarios de este producto. Es por ello que se escriben cada una de las preguntas que corresponden a las variables y seguidamente (en un nivel inferior) se determinan las respuestas correspondiendo estas a la tipología de variable. Si es numérica, se añade una barra donde poder insertar los números. Pero si es categórica, aparece una barra con las diferentes opciones de la variable (coincidiendo con los niveles de la variables descritas en el conjunto de datos).

Para el desarrollo del *BackEnd* se usa un nuevo cuaderno de Jupyter, describiendo las funciones necesarias. En primer lugar se importan las librerías necesarias como NumPy o Flask. El primer paso será la creación del objeto de aplicación que se use para construir el backend y se carga el modelo usando *pickle*. Luego, se procede a definir la página de inicio de la aplicación definiendo una función que represente el archivo html anteriormente descrito. Por último, se define la predicción. Para ello, se usa una función de predicción donde se almacenan los valores de entrada del cuestionario en una matriz que se pasará al modelo directamente para que realice la predicción. Este valor de predicción se almacena en una variable que comparte con el html para luego mostrar este resultado en la aplicación.

Por último, para desplegar la aplicación a la web se usará *PythonAnywhere* donde se suben los archivos html, cuaderno de jupyter (backend) y el modelo de aprendizaje automático para poder subirla a la nube y hacerla accesible a un público general.

B. Repositorio público:

Un repositorio público es un sistema de información que permite tanto el almacenaje de datos y proyectos como un acceso público a este mismo. En el caso de la rama más tecnológica, el repositorio más común es *Github*, un servicio de alojamiento de Internet usado para el desarrollo de *software* y controles de versión usando Git⁹⁵. Este repositorio permite el control de acceso, seguimiento de errores y gestiones de tareas, entre otros. Para facilitar la revisión y consulta del código implementado en el desarrollo de este trabajo, en este proyecto se decide publicar con acceso abierto dicho código. Esta publicación se realiza mediante Github, siguiendo los pasos establecidos por el propio servicio de *Github*.

4. RESULTADOS

4.1. ANÁLISIS DE DATOS

A. Análisis exploratorio y separación entrenamiento-test

Se realiza la carga de los datos en el cuaderno de Jupyter y se comprueba el conjunto de datos, obteniendo un *dataframe* de 450 filas y 131 variables que coinciden con la descripción del conjunto de datos original descrito por los autores (130 variables más una variable “etiqueta”). En este análisis exploratorio se observa que este conjunto de datos está formado por diferentes tipos variables, entre ellas tenemos de ejemplo:

- Variable numérica continua: Variable peso o BMI.
- Variable numérica discontinua: Número de hijos o número de abortos.
- Variable categórica nominal: Nivel de estudios de la madre o cantidad de cigarrillos consumidos durante la adolescencia.

Las variables de tipo categórica nominal se encuentran factorizadas, sustituyendo cada opción de respuesta del tipo test a números. Este cambio facilita el tratamiento de los datos, siendo así todas las variables son de tipo *float* o *int* en Python (Figura 4.1). No obstante, se debe tener en cuenta el tipo de datos original (es decir, el carácter categórico de las variables) en los posteriores análisis.

```
# 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
```

Figura 4.1 Información del *dataframe* original. Imagen descriptiva del conjunto de datos inicial, se observan información sobre columnas, filas, tipo de variables y uso de memoria.

Por otra parte, se revisan las variables con el fin de comprobar que la información que aportan se relaciona con el objetivo de nuestro trabajo. Este conjunto de datos diferencia las variables en cinco tipos: características físicas y mentales de la madre, estilo de vida de la madre, estrato social de la madre, niveles de estrés y focos de estrés de la madre y variables con relación a momento postparto y salud del bebé. El último grupo de variables al ser solamente del momento postparto no se adaptan a la hipótesis de trabajo debido a que no aportan información que ayude a predecir el tipo de parto. Es por ello que se decide eliminar las ocho variables que conforman este último grupo, así se evita problemas de correlación alta.

Teniendo ya el conjunto de datos analizado de manera completa, se procede a la obtención de la variable dependiente (Y) y la matriz de variables independientes (X). Como se ha indicado en el apartado de metodología, se estudia la distribución de los valores de la variable Y comprobando que solamente 81 de las 450 observaciones corresponden a partos prematuros.

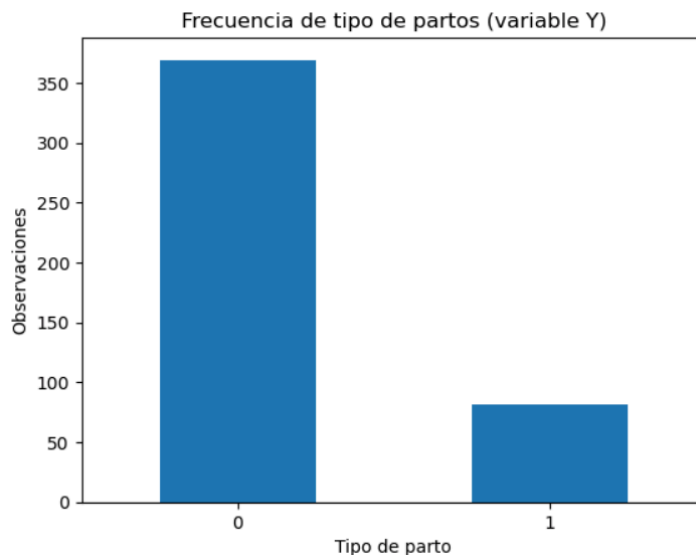


Figura 4.2 Frecuencia tipo de partos. Gráfica descriptiva de la distribución de los valores de la variable independiente. El número 0 corresponde a los partos naturales y el grupo 1 a los partos prematuros.

Como se puede ver, los datos para la variable Y están desequilibrados siendo mayoritarios los datos de partos naturales. Este factor se tiene en cuenta en los siguientes pasos del análisis ya que afecta a los modelos de predicción pudiendo llevar a obtener resultados engañosos. Se puede obtener un valor de predicción alto, debido a la alta probabilidad de obtener un resultado de parto natural en comparación con parto prematuro. Para este estudio se decide no igualar las proporciones y mantener estas en todos los pasos.

Por último, se separan los datos en conjunto de entrenamiento y test. Se escoge un tamaño del 20% para los datos de tipo test que nos da como resultado 90 observaciones para test y 360 observaciones para entrenamiento. Se obtienen como producto final X_{train} , X_{test} , y_{train} y y_{test} . Viendo la distribución desigual de la variable independiente, se añade la opción de “*stratify*” para mantener la proporción de partos prematuros y no prematuros en los conjuntos de datos y_{train} y y_{test} .

B. Preprocesamiento: Estudio de valores nulos

El primer paso dentro del estudio de valores nulos es comprobar la existencia de estos dentro del conjunto de datos. En este caso, vemos que existen valores nulos en X_{train} y X_{test} , pero en y_{train} y y_{test} no se encuentra ninguno. En concreto, los valores nulos se encuentran en 22 variables (columnas) siendo estas de diferente tipo (3 de tipo numéricas y 19 de tipo categórica).

Habiendo identificado el tipo de variable, se estudia la distribución de las variables numéricas para conocer si son de tipo numérica continua o discontinua ya que implica un tratamiento diferente. Solamente una variable (*weight before delivery*) de estas tres es de tipo continua y por tanto, para el tratamiento de valores nulos se usa la sustitución de ellos por el valor de la media. Por otro lado, en las otras dos variables (*No of sibling*, *Miscarriage History*) se opta por la opción de sustituir los valores nulos por los valores más comunes de cada variable.

Para las 19 variables restantes, al ser de tipo categórica se opta por aplicar tres métodos diferentes obteniendo tres conjuntos de datos. Estas tres opciones son:

- Creación de una nueva clase, denominada con el número 10 para evitar problemas de lectura de datos como pueden dar las variables tipo *string*. Se obtiene el conjunto de datos ***X_train1 / X_test1***.
- Sustitución por el valor más común de cada una de las variables. Se obtiene el conjunto de datos ***X_train2 / X_test2***.
- Aplicación del algoritmo kNN vecinos para la imputación de los datos. Se obtiene el conjunto de datos ***X_train3 / X_test3***.

En los tres casos, al aplicar estos modelos se obtienen tres conjuntos de datos sin valores nulos. En el caso del tratamiento de valores nulos para variables continuas, se aplica directamente en el conjunto de datos original y se sustituyen los valores modificando este mismo conjunto de datos. En el caso de las variables categóricas, al no tener una opción totalmente válida ni conociendo una manera clara de poder decidir qué método es el más adecuado, se ha preferido optar por la aplicación de tres métodos diferentes y separar cada uno de estos resultados en un conjunto de datos diferente para posteriormente comparar cuál de las opciones es más adecuada.

C. Preprocesamiento: Estudio de valores atípicos

El último paso del preprocesamiento es la detección y tratamiento de valores atípicos. Debido al alto número de variables y a los diferentes tipos de variables que se manejan en este conjunto de datos, se opta por el uso de algoritmos como PCA y *Isolation Tree*. Se descartan las opciones más comunes como la representación mediante diagrama de cajas (estudio de cuartiles) o por histograma (estudio de desviación estándar) debido al número de datos que se manejan complicando la comparación y comprobación de cada una de las variables.

En los dos algoritmos que se aplican, se siguen tres pasos fundamentales para la detección de valores atípicos:

- Desarrollo del modelo: En este paso se describen las características más técnicas del modelo como el porcentaje de contaminación permitido (10%) en la técnica de PCA o la contaminación y número máximo de muestras en *Isolation Forest*. Se aplica cada algoritmo a los tres conjuntos de datos y se calculan los niveles y puntuaciones de predicción.

Para comprobar la correcta aplicación del algoritmo, se muestran los parámetros de cada modelo y los valores de varianza explicada por cada componente principal en el caso de PCA. En el caso de los parámetros se puede comprobar el grado de contaminación escogido y la estandarización de los datos automáticamente. Se observa que en los tres conjuntos de datos, la varianza explicada por los primeros componentes es un valor pequeño (oscila entre 18-32%).

Para la técnica de *Isolation Forest*, se define el modelo tanto para el conjunto de datos de entrenamiento como de test. En este caso, también obtenemos los parámetros pudiendo comprobar el grado de contaminación que se ha descrito (10%), como el número máximo de muestras que se extraen del conjunto de entrenamiento (40), el número de árboles en el conjunto (100) y el número de

trabajos que se ejecutan en paralelo. Luego, se muestra en forma numérica y gráfica la importancia de las variables siendo en los tres casos valores bajos (del orden de 10^{-2}).

- Cálculo del treshold: Como el enfoque del tratamiento de valores nulos es la limitación de estos, se implementa un paso que calcule el límite que se usará para determinar qué valores se consideran valores atípicos y cuáles no. Tanto en PCA como en *Isolation Forest*, se calcula mediante una función el valor numérico del límite y también se muestra en histograma la distribución de la puntuación de valores atípicos. Este gráfico puede ayudar a discernir si el valor calculado automáticamente por el algoritmo es correcto o si se prefiere explorar otros valores alternativos.

En la opción de PCA se obtienen valores límite muy altos del orden de 10^{35} y en el caso de *Isolation Forest* corresponden valores límite muy pequeños (del orden de 10^{-17}). Además, observando las gráficas no corresponden los valores numéricos con los observados en la gráfica indicando que el rango de puntuaciones de valores atípicos en los dos casos son muy altos imposibilitando así la obtención de estos valores límite de manera gráfica.

La diferencia de ordenes entre un método y otro se explica por los enfoques que tienen cada uno de los algoritmos para estudiar los valores atípicos. En el caso de PCA, al usar el valor de error de reconstrucción, se obtienen valores muy grandes correspondientes a los puntos más alejados del promedio.

Por el contrario, para el método de *Isolation Forest* se mide la distancia entre variables más cercanas a otras dando valores de orden más bajo.

- Determinación de grupos normal y outlier: Con el límite establecido y el algoritmo aplicado a cada conjunto de datos, se muestra mediante una tabla la cantidad de valores atípicos que detecta cada algoritmo. Para los dos algoritmos en los tres conjuntos de datos se detectan 36 valores atípicos y 324 valores “normales”.

Además, se calcula una matriz de confusión para comprobar la predicción de los valores atípicos pudiendo distinguir entre falsos positivos (valores detectados como *outliers* pero que no lo son) o falsos negativos (valores no considerados como *outliers* pero si lo son). Para los conjuntos de datos X_{train2} y X_{train3} , los dos algoritmos detectan solamente 6 valores como *True Positive (TP)* (auténticos valores atípicos). Por otro lado, en el conjunto de datos X_{train1} se obtienen 5 valores TP con el modelo de PCA y 8 valores TP en el modelo de *Isolation Forest*.

Detección valores atípicos: matriz de confusión

(A)	(B)	(C)																																				
<table> <tr> <th>Pred</th><th>0</th><th>1</th></tr> <tr> <th>Actual</th><td></td><td></td></tr> <tr> <th>0</th><td>264</td><td>31</td></tr> <tr> <th>1</th><td>60</td><td>5</td></tr> </table>	Pred	0	1	Actual			0	264	31	1	60	5	<table> <tr> <th>Pred</th><th>0</th><th>1</th></tr> <tr> <th>Actual</th><td></td><td></td></tr> <tr> <th>0</th><td>265</td><td>30</td></tr> <tr> <th>1</th><td>59</td><td>6</td></tr> </table>	Pred	0	1	Actual			0	265	30	1	59	6	<table> <tr> <th>Pred</th><th>0</th><th>1</th></tr> <tr> <th>Actual</th><td></td><td></td></tr> <tr> <th>0</th><td>265</td><td>30</td></tr> <tr> <th>1</th><td>59</td><td>6</td></tr> </table>	Pred	0	1	Actual			0	265	30	1	59	6
Pred	0	1																																				
Actual																																						
0	264	31																																				
1	60	5																																				
Pred	0	1																																				
Actual																																						
0	265	30																																				
1	59	6																																				
Pred	0	1																																				
Actual																																						
0	265	30																																				
1	59	6																																				
(D)	(E)	(F)																																				
<table> <tr> <th>Pred</th><th>0</th><th>1</th></tr> <tr> <th>Actual</th><td></td><td></td></tr> <tr> <th>0</th><td>267</td><td>28</td></tr> <tr> <th>1</th><td>57</td><td>8</td></tr> </table>	Pred	0	1	Actual			0	267	28	1	57	8	<table> <tr> <th>Pred</th><th>0</th><th>1</th></tr> <tr> <th>Actual</th><td></td><td></td></tr> <tr> <th>0</th><td>265</td><td>30</td></tr> <tr> <th>1</th><td>59</td><td>6</td></tr> </table>	Pred	0	1	Actual			0	265	30	1	59	6	<table> <tr> <th>Pred</th><th>0</th><th>1</th></tr> <tr> <th>Actual</th><td></td><td></td></tr> <tr> <th>0</th><td>265</td><td>30</td></tr> <tr> <th>1</th><td>59</td><td>6</td></tr> </table>	Pred	0	1	Actual			0	265	30	1	59	6
Pred	0	1																																				
Actual																																						
0	267	28																																				
1	57	8																																				
Pred	0	1																																				
Actual																																						
0	265	30																																				
1	59	6																																				
Pred	0	1																																				
Actual																																						
0	265	30																																				
1	59	6																																				

Figura 4.3 Matrices de confusión en los métodos de detección de valores atípicos.

(A) Matriz de confusión correspondiente al método PCA para el conjunto de datos X_{train1} . (B) Matriz de confusión correspondiente al método PCA para el conjunto de datos X_{train2} . (C) Matriz de confusión correspondiente al método PCA para el conjunto de datos X_{train3} . (D) Matriz de confusión correspondiente al método *Isolation Forest* para el conjunto de datos X_{train1} . (E) Matriz de confusión correspondiente al método *Isolation Forest* para el conjunto de datos X_{train1} . (F) Matriz de confusión correspondiente al método *Isolation Forest* para el conjunto de datos X_{train1} .

La matrices de confusión también nos indican que los dos métodos no son muy precisos ya que la cantidad de valores TP y TN no concuerdan con los predichos de antemano por el modelo (de 36 valores atípicos, solo detecta 6 - 8 como verdaderos valores atípicos). No obstante, comparamos los modelos para comprobar si entre ellos existe alguna diferencia con respecto a la capacidad de detección.

Para esta comparación se representa en una tabla los valores límite de cada modelo y se muestra la predicción de cada uno de ellos con respecto a unos valores aleatorios del conjunto de datos. Se observa que los dos modelos fallan en alguna de las predicciones, demostrando que no son modelos muy optimizados. Para comprobar de manera numérica la capacidad de predicción de estos dos modelos, se calcula el valor de precisión obteniendo como resultado valores que oscilan en el 75% para los 6 casos. Aunque este resultado de precisión muestre un valor bastante aceptable, se decide no eliminar ningún valor atípico ya que representan una población muy pequeña con respecto al total del conjunto de datos (menos del 10%).

D. Reducción de dimensionalidad

Por último, antes de proceder al estudio de predicción mediante algoritmos de aprendizaje automático, se realiza el paso de reducción de dimensionalidad. Se decide utilizar la estrategia de selección de variables mediante un método integrado como *Random Forest Classifier* debido a las características del conjunto de datos (variables categóricas y numéricas) y por cumplir el objetivo de la obtención de las variables más representativas de este conjunto de datos para la definición de un cuestionario. Para ello, aplicamos este algoritmo a cada uno de los conjuntos de datos y mostramos en una tabla los valores de las diez variables más importantes según la clasificación realizada por el modelo.

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

Figura 4.4 Conjunto de tablas de relevancia de variables. (A) Tabla con resultados de relevancia de variables mediante *Isolation Forest* del conjunto de datos $X_{train1N}$. (B) Tabla con resultados de relevancia de variables mediante *Isolation Forest* del conjunto de datos $X_{train2N}$. (C) Tabla con resultados de relevancia de variables mediante *Isolation Forest* del conjunto de datos $X_{train3N}$.

Para los tres casos, se realiza un gráfico mostrando la importancia de las variables en cada uno de los modelos, viendo en todos los casos que a partir de la séptima variable disminuye el valor de importancia más de la mitad del valor inicial.

Se crean tres nuevos conjuntos de datos ($X_{train1d}$, $X_{train2d}$, $X_{train3d}$) seleccionando solamente las diez variables extraídas mediante este modelo. Se destaca que los tres modelos coinciden en **9 de las 10** variables más importantes: nivel de hemoglobina, edad de la madre, edad del padre, peso antes del embarazo, altura, BMI, años de casados, problemas gástricos durante el embarazo y nivel educación de la madre. La otra variable que se diferencia es número de bebés (embarazo múltiple o no) que aparece en dos de los conjuntos de datos o ingresos anuales de la pareja.

Esto es un indicativo que a pesar de haber aplicado diferentes métodos con respecto al tratamiento de valores nulos, se mantienen las características más relevantes del conjunto de datos original. Siendo así que la influencia de las variables más importantes se mantiene en estos tres conjuntos de datos. Este hecho, facilitará la aplicación y comparación de los modelos de aprendizaje automático que se aplican en el siguiente apartado.

4.2. APRENDIZAJE AUTOMÁTICO - DEFINICIÓN MODELOS

A. Árboles de decisión

Para los tres conjuntos de datos, creamos el modelo de árbol de decisión dejando los hiperparámetros que vienen por defecto. Se entrena el modelo y se calcula la predicción para nuestros datos tipo test. Para la valoración de este modelo se usan los valores f1 y AUC, pudiendo así conocer si se llega al valor umbral definido en los objetivos iniciales del trabajo.

	$X_{train1d}$	$X_{train2d}$	$X_{train3d}$
Valor f1	0.7972	0.7262	0.6986
Valor AUC	0.5979	0.5194	0.5346

Tabla 4.1 Valores f1 y AUC. Valores f1 y AUC obtenidos en el modelo de árboles de decisión para los tres conjuntos de datos.

Como se puede observar, para los valores obtenidos de f1 solamente el conjunto de datos $X_{train1d}$ supera el valor umbral preestablecido (75%). Los otros dos valores correspondientes a $X_{train2d}$ y $X_{train3d}$ son valores inferiores a este umbral, indicando que se debe optimizar este modelo. Para el valor AUC ninguno de los resultados son óptimos con respecto al valor umbral establecido (son inferiores al 60%).

Se concluye que este modelo debe repetirse optimizando su ejecución con respecto a los conjuntos de datos que se manejan. Con estos resultados se aprecian las diferencias entre la medición que realiza f1 y el valor AUC. Mientras que AUC es el área bajo la curva ROC calculada en los umbrales entre la tasa de verdaderos positivos y tasa de falsos positivos, F1 es un cálculo directo que involucra la recuperación y precisión general del modelo. Es por ello, que seguramente el valor de AUC esté influenciado por el desequilibrio del conjunto en sí afectando ello a la capacidad de obtención falsos positivos y verdaderos positivos.

B. Bosques aleatorios

El primer paso antes del desarrollo del modelo es la conversión de las variables de carácter categórico a variables tipo *dummy*. Para el conjunto de datos *X_train1d* y *X_train3d* se transforman las variables “*Education*” y “*Gastric_preg*” obteniendo en total 22 columnas ya que estas dos variables tienen 6 niveles. Por otro lado, para el conjunto de datos *X_train2d*, se transforman las variables “*Education*”, “*Gastric_preg*” y “*Family_Income*” obteniendo un total de 25 columnas. Con estos cambios se crean nuevos conjuntos de datos, siendo *ohe_data1*, *ohe_data2* y *ohe_data3*.

Teniendo las variables transformadas, se procede a la definición y aplicación del modelo a los tres conjuntos de datos. Para la comprobación de la capacidad de predicción de los modelos obtenidos, se calcula el valor de f1 y AUC.

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

Tabla 4.2 Valores f1 y AUC. Valores f1 y AUC obtenidos en el modelo de bosques aleatorios para los tres conjuntos de datos.

En los tres casos se obtienen valores inferiores a los requeridos para la confirmación del modelo como óptimo, estando el valor f1 para los tres casos casi superando este umbral (75%). Este resultado es un indicativo que los modelos definidos funcionan bien pero solamente se deben redefinir los hiperparámetros para considerarlos óptimos. Por otro lado, el valor AUC que se obtienen son los más bajos (el rango para los valores AUC va del 0.5 a 1) indicando que el rango de falsos positivos y verdaderos positivos está muy desajustado en este modelo.

C. Máquinas de vectores de soporte

Antes de empezar la definición de este modelo, se escalan los datos creando así tres conjuntos de datos nuevos donde todas las variables tienen un valor mínimo de 0 y un valor máximo de 1. Estos nuevos conjuntos de datos son: *df_scaled1*, *df_scaled2* y *df_scaled3*.

El siguiente paso es el cálculo de los hiperparámetros más adecuados para cada nuevo conjunto de datos. Se da una combinación de cada opción de los hiperparámetros aplicando cada una de ellas a los conjuntos de datos pudiendo así comparar luego con

cuál se ha obtenido un resultado más óptimo. Se muestra el valor de los parámetros y se calculan los valores de precisión, robustez, valor f1 y la precisión del modelo con mejor resultado.

Se puede observar que los valores de precisión, robustez y f1 vienen separados dependiendo del valor de la variable dependiente (Y). En este caso, se ve el impacto del desequilibrio de la variable Y, ya que para la clase 0 (partos naturales) se obtienen valores bastante elevados (superando el 0.85 en los tres parámetros). No obstante, en la clase 1 (partos prematuros) los valores de estos tres parámetros son muy bajos llegando incluso a cero.

		df_scaled1	df_scaled2	df_scaled3
Hiperparámetros	C	10	0.1	10
	Gamma	0.1	1	0.1
	Kernel	rbf	linear	rbf
Valores de evaluación de hiperparámetro - clase 0 de variable Y	Valor precisión	0.85	0.82	0.85
	Valor robustez	0.99	1	0.99
	Valor f1	0.91	0.90	0.91
Valores de evaluación de hiperparámetro - clase 1 de variable Y	Valor precisión	0.75	0	0.75
	Valor robustez	0.19	0	0.19
	Valor f1	0.30	0	0.30

Tabla 4.3 Valores hiperparámetros y medidas de evaluación. Se muestran los valores de hiperparámetros para cada conjunto de datos junto con los resultados de evaluación de cada uno de los modelos.

Conociendo ya que hiperparámetros funcionan mejor para cada conjunto de datos, procedemos a la definición de los tres modelos para cada uno de los conjuntos de datos. Se obtiene el resultado de predicción usando los valores AUC y f1.

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

Tabla 4.4 Valores f1 y AUC. Valores f1 y AUC obtenidos en el modelo de máquinas de soporte para los tres conjuntos de datos.

En este caso, en los conjuntos de datos 1 y 3 se obtienen valores de f1 que superan el umbral previsto para los modelos de predicción óptimo. No obstante, el segundo modelo no llega ni a alcanzar el 0.70 del valor de f1. Esto se puede deber a causa de las variables que definen cada modelo ya que los conjuntos de datos *df_scaled1* y *df_scaled3* tienen las 10 mismas variables. Justamente los dos modelos de SVM de estos dos conjuntos de datos se han definido con los mismos valores de hiperparámetros (C = 10, gamma = 0.1 y

kernel = rbf). No obstante, se ven diferencias con respecto a los valores de f1 y AUC ya que las observaciones no son las mismas (proceden de datos diferentes debido a los diferentes tratamientos de valores nulos).

Estos modelos no se optimizarán ya que no se pueden ajustar mejor los hiperparámetros que los definen, siendo así que estos modelos utilizados son los definitivos con respecto al algoritmo de máquinas de vectores de soporte.

D. Redes neuronales artificiales

Para la implementación del algoritmo de redes neuronales, se usarán los mismos conjuntos de datos estandarizados y con las variables tipo *dummy*: *df_scaled1*, *df_scaled2* y *df_scaled3*. Ya con los conjuntos de datos preparados, se procede a la definición del modelo junto con sus características (número de capas, función de activación y función de coste).

Debido a las características de la variable independiente (clases desequilibradas) y con el fin de optimizar el funcionamiento de la red neuronal, se realizan varios pasos:

- Añadir una capa “*dropout*”
- Añadir función “*early stopping*”
- Cálculo de bias adecuado
- Cálculo de pesos inicial

Después de definir el modelo, se realiza el cálculo del valor bias inicial para intentar mejorar el ajuste al conjunto de datos. Se calcula el valor de *Loss* tanto para el valor inicial como para el valor con el *bias* corregido y así para comparar si el ajuste de bias es necesario o no.

	df_scaled1	df_scaled2	df_scaled3
Valor Loss inicial	0.4822	0.6714	0.8258
Valor Loss final	0.4747	0.4737	0.4742

Tabla 4.5 Valores Loss. Valores *Loss* obtenidos en el paso de cálculo de *bias* como método de comparación entre los dos modelos.

Como se puede observar, los valores *loss* disminuyen en todos los casos indicando que el control del valor *bias* influye de manera positiva en la definición del modelo. Es por eso, que en todos los modelos se aplica el nuevo valor *bias* en el modelo definitivo.

Seguidamente se realiza una comprobación gráfica de los valores *Loss* pudiendo ver los valores se encuentran a un punto de distancia (comparación entre *Loss* inicial y *Loss* con *bias* corregido). Además, siguen la misma tendencia descendente ajustándose los valores de entrenamiento con los valores de validación.

Por último, se calculan los valores de peso para la variable Y pudiendo así otorgarle más peso a la clase 1 (parto prematuro) e intentar igualar las dos clases para evitar un sobreajuste sobre la clase 0. Con el peso ajustado, se procede a la definición de este modelo con todas las características descritas.

Para poder comprobar la realización del modelo con respecto a la predicción de los valores, se realizan los gráficos para los valores *Loss*, *precisión* y *robustez*. Por último, se calcula de manera numérica los valores f1 y auc mostrados.

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

Tabla 4.6 Valores f1 y AUC. Valores f1 y AUC obtenidos en el modelo de máquinas de soporte para los tres conjuntos de datos.

Se obtienen los mismos valores para los 3 conjuntos de datos siendo estos valores bastante bajos. Se comprueba que en las redes neuronales no son capaces de discernir las diferencias entre los tres conjuntos de datos, quedándose solamente con las características más generales. Es así que no existe ninguna distinción entre estos conjuntos de datos.

Por otra parte, los valores tan bajos de f1 y AUC indican que el algoritmo de redes neuronales artificiales no se adapta bien al problema al que se tiene que enfrentar en este trabajo. Por tanto, este modelo queda directamente descartado ya que no se pueden optimizar más este modelo.

4.3. APRENDIZAJE AUTOMÁTICO - OPTIMIZACIÓN DE LOS MODELOS

Para los dos modelos de aprendizaje automático donde no se han alcanzado los valores de f1 superiores a 75%, se repite la ejecución de estos mismos modelos simplemente modificando algunos de los hiperparámetros. En el caso de las máquinas de vectores de soporte y las redes neuronales artificiales, no se pueden optimizar y es por ello que no se añaden en este apartado.

A. Árboles de decisión

Como se comenta en el apartado de metodología, existen varios hiperparámetros que se pueden controlar en el algoritmo de árboles de decisión. En este caso, se opta por la definición de la profundidad máxima del árbol definiendo este valor en 6. Ahora, el nuevo modelo nos da como resultado los valores mostrados en la tabla 3.7.

	X_train1d	X_train2d	X_train3d
Valor f1	0.8140	0.7755	0.7420
Valor AUC	0.6115	0.5489	0.5

Tabla 4.7 Valores f1 y AUC. Valores f1 y AUC obtenidos en el modelo de árboles de decisión para los tres conjuntos de datos.

En estos resultados podemos ver el efecto de cómo la definición de diferentes parámetros, nos permite ajustar el modelo hasta la obtención de los valores dentro del rango estipulado. Los dos conjuntos de datos *X_train1d* y *X_train2d* obtienen valores superiores al 75%, haciendo de ellos modelos óptimos para su uso en la predicción de partos prematuros. No obstante, a pesar del ajuste de los hiperparámetros, el modelo *X_train3d* no llega al umbral, siendo directamente descartado para el objetivo de este trabajo.

B. Bosques aleatorios

Debido a la similitud del actual algoritmo con respecto al algoritmo anterior, se decide aplicar la misma estrategia de optimización modificando solamente el valor de profundidad y definiéndolo con un valor de 6. Se decide solamente modificar este hiperparámetro porque se prefiere ir ajustando poco a poco los modelos pudiendo controlar a qué son debido los cambios, ante la opción de modificar sustancialmente el modelo sin poder conocer la causa de la mejoría de ajuste. También se vuelven a calcular los valores de f1 y AUC para los tres conjuntos de datos.

	ohe_data1	ohe_data2	ohe_data3
Valor f1	0.8036	0.7681	0.7681
Valor AUC	0.5869	0.5422	0.5422

Tabla 4.8 Valores f1 y AUC. Valores f1 y AUC obtenidos en el modelo de redes neuronales artificiales para los tres conjuntos de datos.

Como se puede observar, los valores de f1 mejoran en los tres modelos superando así el valor umbral del 75%. Destaca de estos datos, la similitud de los resultados para los conjuntos de datos *ohe_data2* y *ohe_data3*, habiendo ocurrido en el modelo sin optimizar. Esta similitud puede indicar que entre los dos conjuntos de datos no existen diferencias muy relevantes, haciendo que el modelo *Random Forest* no sea capaz de distinguir entre uno u otro.

Para finalizar el apartado de modelos de aprendizaje automático, se realiza una tabla donde aparezcan solamente los valores de f1 para los tres conjuntos de datos en los 4 algoritmos que se han aplicado. El valor de f1 más alto de esta tabla será el modelo de predicción aceptado en este trabajo y el cual se aplicará para el posterior paso de desarrollo de la aplicación web.

Se descarta el uso del valor AUC para la decisión de modelo ya que como se ha visto a lo largo de este apartado de resultados, permanece en un valor bastante bajo mostrando así una capacidad pobre de ajuste respecto a este tipo de datos. A pesar de ser un debate de actualidad, en este trabajo se demuestra que el valor AUC sí se ve afectado si la variable dependiente está desequilibrada.

	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

Tabla 4.9 Valores f1 final de todos los modelos. Valores f1 obtenido en los diferentes modelos de aprendizaje automático para los tres conjuntos de datos.

En esta tabla podemos ver de manera resumida el funcionamiento de los modelos predictivos con respecto a la variable de calidad f1. Destacan los valores tan bajos para el modelo de redes neuronales artificiales, indicando que no son modelos que sirvan para la predicción de esta variable dependiente. Por otro lado, se ve que el conjunto de datos 2 - máquinas de vectores de soporte y conjunto de datos 3 - árboles de decisión se quedan justo en el límite o con valores relativamente cercanos, siendo esto un ejemplo que la decisión del tratamiento de valores atípicos si puede influir en los posteriores pasos del procedimiento de aprendizaje automático (llegando a limitarlo para la función de predicción).

Por último, destacar que 7 de los 12 valores f1 superan el umbral representando más de la mitad de los modelos calculados. De estos siete modelos, el **conjunto de datos 3 - árboles de decisión** es el que se define con el valor más alto de f1, siendo por eso el modelo escogido para el desarrollo web.

4.4. APLICACIÓN WEB

El primer paso para el desarrollo web y la aplicación del modelo predictivo a esta misma consiste en la obtención del modelo en un formato legible para la web. Para ello se aplica la librería *pickle* que permite obteniendo como resultado un archivo en el repositorio indicado.

Para la creación de la página en sí y el estilo, se desarrolla un código en html. En este código, se describen las 10 variables del modelo de manera que se escribe la pregunta a la que corresponde la variable y se añaden las posibles opciones. Para que el modelo pueda entender cada respuesta, se adjudica a cada opción escrita un número que correspondiente a la codificación de las diferentes respuestas del modelo. En el caso de las variables numéricas, se facilita una barra donde se puedan añadir los números (habiendo indicado que la variable es de tipo número. Para evitar pequeñas erratas, se marca como incorrecto los números iguales o inferiores a 0.

Aplicación web

(A)

```
<div class="mb-3">
  Nivel educación de la madre
  <select class="form-select" aria-label="Default select example">
    <option value="0">Sin estudios </option>
    <option value="1">Estudios básicos </option>
    <option value="2">Estudios obligatorios (secundaria) </option>
    <option value="3">Estudios Bachiller / Formación profesional </option>
    <option value="4">Estudios Grado Universitario </option>
    <option value="5">Estudios Máster </option>
    <option value="6">Estudios Doctorado </option>
  </select>
</div>
```

(B)

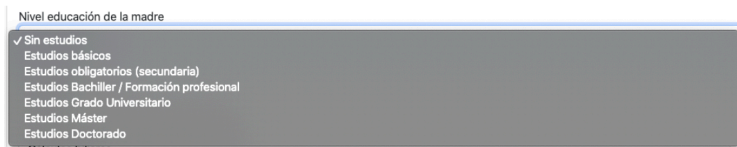


Figura 4.5 Comparación código html y página web. (A) Código html correspondiente a la descripción de opciones para la variable “*Education*”. (B) Visualización del cuestionario con las preguntas y respuestas en el archivo de html finalizado.

Para evitar la creación de archivos adjuntos de tipo css o javascript para la definición del estilo de la página, se decide implementar *Bootstrap* directamente en el archivo de html e ir definiendo los estilos como tipo de botón, cabecera, etc.

Teniendo el estilo y la web definida, se desarrolla el código correspondiente al *Backend* de la página web. Con este *Backend* definido, se establece la conexión entre las opciones escogidas por el usuario en el cuestionario que se muestra y el modelo de predicción descrito en el archivo tipo plk.

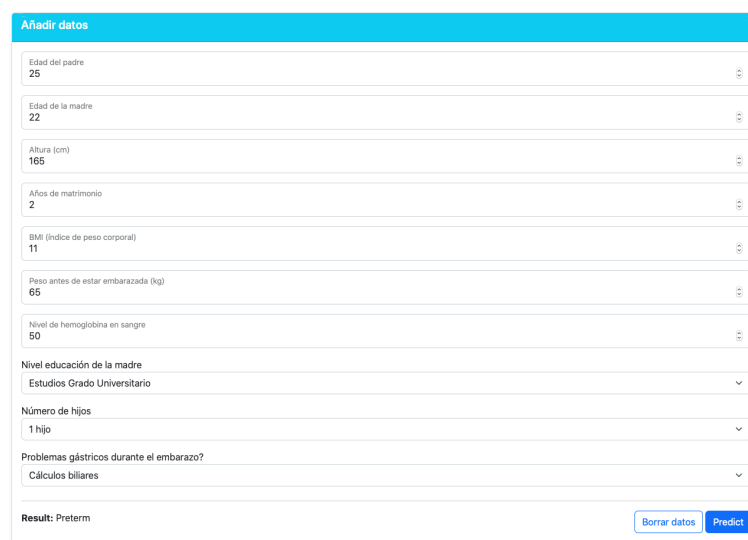


Figura 4.6 Resultado página web. Imagen de la página web con el cuestionario y el resultado obtenido en la predicción.

Para finalizar, se suben los archivos correspondientes de este desarrollo web a la aplicación *PythonAnywhere* pudiendo subir así el modelo de machine learning y la web a la nube.

5. CONCLUSIONES Y TRABAJOS FUTUROS

5.1. Conclusiones

Antes de empezar con el desarrollo de este apartado, solamente informar que las conclusiones van a ser comentadas con respecto a los objetivos generales descritos en el apartado 1.2 de esta misma memoria. Así se confirma el cumplimiento de todos los objetivos descritos en este trabajo obteniendo tanto el perfil multifactorial para el estudio de partos prematuros como los modelos de aprendizaje automático y la aplicación web.

Con respecto al primer objetivo de la determinación del perfil multifactorial específico para la predicción de partos prematuros, se han obtenido tres conjuntos de datos reducidos que comparten nueve de las diez variables escogidas. A pesar de tener tres modelos con diferentes tratamientos de valores nulos, los tres coinciden prácticamente en la determinación de la importancia de las variables siendo indicativo de la gran influencia de estas variables dentro de la determinación de parto prematuro o no.

Por otro lado, 6 de las 10 variables extraídas corresponden a características físicas de la madre (como el peso o la altura); no obstante, también aparecen variables de carácter personal como el nivel de educación de la madre o los años de casados. Este resultado demuestra la importancia de la realización de estudios multifactoriales para el abordaje de los partos prematuros. Se demuestra que no sólo las características médicas de la madre pueden influenciar en el tipo de parto. Como bien se ha ido indicando a lo largo de esta memoria (apartado 1, apartado 2), se quiere poner en el foco la influencia de las desigualdades socio-económicas como factor de riesgo en los partos prematuros. En este estudio, se ha visto que el nivel de educación de la madre influye estando este ligado a las condiciones económicas de la madre. También en uno de los conjuntos de datos aparece como factor los ingresos económicos de la pareja. Esto refuerza una de las hipótesis de trabajo inicial, demostrando que el factor socio-económico sí juega un papel importante dentro de la determinación de riesgo con respecto al parto prematuro.

Así pues, los resultados obtenidos siguen la línea de estudio que se ha planteado en este trabajo, pero sorprende la importancia de la variable años de casados de la pareja. No es una variable que se considera en ningún estudio relevante, pero se puede postular que esta variable está relacionada con el nivel de estrés de la madre.

Además, aparecen otros factores que ya se tenían en cuenta como la edad de los progenitores o el peso de la madre. Con estos resultados, la implementación de este nuevo modelo con carácter predictivo de partos prematuros no supone un problema ya que solamente implica la obtención de nuevas variables de fácil obtención como la altura de la madre o los años de casados combinándolas con variables que ya se toman de manera rutinaria.

Y por otro lado, con respecto al segundo objetivo se obtiene que el modelo con mejores datos de predicción es el conjunto de datos tipo 3 con el algoritmo de árboles de decisión. Este resultado no es sorprendente ya que uno de los algoritmos que más se usan para la predicción son los algoritmos de tipo árbol (árboles de decisión o bosques aleatorios) siendo estos los que mejores resultados dan. En este trabajo se ha querido explorar diferentes tipos de algoritmos, aún sabiendo que, por ejemplo, las redes neuronales no son un algoritmo muy optimizado para este objetivo. Así se ha visto que los resultados de las redes neuronales dan un valor de f1 bastante bajo indicando una capacidad baja de predicción.

En este resultado también se demuestra que el método de detección y tratamiento de valores nulos más eficiente es la imputación de estos mismos valores mediante el algoritmo kNN-vecinos. De las 19 variables (de tipo categórica) con valores nulos había más de 200 observaciones con estas características. Es por ello que se puede explicar que este método de imputación de mejores resultados ya que implica la sustitución de estos valores por los valores más probables de más de la mitad de la población del estudio. Por contra, los otros dos métodos ofrecen una solución más rápida, pero en este caso se queda incompleta. La creación de una nueva clase como la sustitución por el valor más común modifican el comportamiento de estas variables, siendo luego un impedimento para los modelos de aprendizaje automático.

5.2. Seguimiento de la planificación

La planificación de este trabajo se ha basado en el cumplimiento de los objetivos generales y específicos. Para ello, se han ido perfilando unos diagramas de Gantt con respecto al desarrollo de este trabajo cumpliendo objetivos estipulados en un tiempo concreto.

Con respecto a la planificación referida a los pasos dados para el cumplimiento de los objetivos, sí que se ha realizado conforme se describió al inicio del desarrollo de este proyecto. Se han seguido todos los pasos indicados en la planificación, tanto los pasos de análisis de datos como la descripción de los modelos de aprendizaje automático y su optimización.

No obstante, se han tenido que cambiar y repetir varios pasos para alcanzar los objetivos, impidiendo en estos casos el cumplimiento de la planificación con respecto al tiempo de desarrollo. En concreto, el análisis exploratorio se ha tenido que realizar de manera más exhaustiva en la etapa central de desarrollo del proyecto debido a los problemas de correlación que dieron varias variables en los últimos análisis. Este hecho, nos hizo percatarnos de que existían variables post parto que estaban muy relacionadas con la variable dependiente.

Tampoco se tuvo en cuenta el paso del estudio de balanceo de la variable dependiente, teniendo que implantarlo y realizar un pequeño estudio sobre el impacto y ajustes en los modelos de aprendizaje automático.

En los apartados de estudio de valores nulos y valores atípicos, no se consideró la dificultad del conjunto de datos que se iba a manejar. Con respecto a los valores nulos, se

tiene un gran número de estos impidiendo el uso de la estrategia de la eliminación. Además, al tener diferentes tipos de variables, se realizó un estudio sobre las diferentes estrategias para el tratamiento de valores nulos para cada uno de estos. Todo esto impuso una modificación de la planificación temporal, teniendo que invertir más tiempo en este paso.

Con el estudio de valores atípicos, no se tuvo en cuenta el tiempo necesario para la valoración y conocimiento de los métodos a implantar. A causa de la cantidad de variables a estudiar y la tipología de estas (numéricas y categóricas), se descartaron las opciones más comunes como el estudio por cuartiles o desviación estándar. Por tanto, se tuvo que dedicar un tiempo no planificado al estudio de estas opciones y la implementación en el conjunto de datos.

En el caso de la reducción de dimensionalidad, se tuvo que cambiar el método propuesto inicialmente (PCA). Este cambio se debe a la dificultad técnica para la implementación debido a que al tener variables de tipo categórica también se debía separar el conjunto de datos y aplicar en estas otra técnica (MCA). Además, como se explica en el apartado de metodología, se decide optar por métodos de selección de variables. El paso de reducción de dimensionalidad no solo se realiza para mejorar la predicción de los modelos, sino también para la definición de un cuestionario que se muestre en la página web. Es así, que se necesita que no se alteren las variables iniciales. Por ello, se descarta el uso de métodos como PCA o MCA con los que se reduce la dimensión mediante la alteración de las variables en sí, siendo imposible luego rescatar los valores de estas variables.

Por último, con respecto a la implementación de los algoritmos de aprendizaje automático, se ha seguido la planificación requerida. Solamente, se ha adaptado el tiempo para el estudio de requisitos de los modelos con respecto al conjunto de datos y la implementación de estos modelos. Siendo el mismo ajuste temporal necesario en el paso del desarrollo de la aplicación web.

Como se puede ver, debido a todas las modificaciones y tiempo extra que se ha tenido que invertir en el estudio y corrección de métodos del apartado de análisis de datos, la planificación temporal se ha ido modificando con respecto a la propuesta inicialmente. No obstante, todos los problemas que han supuesto, se han ido solucionando ajustando el tiempo de otros apartados y añadiendo horas extras que no se consideraban para la realización de este trabajo.

5.3. Impacto ético-sociales, sostenibilidad y diversidad

En este trabajo, se han conseguido realizar todos los impactos positivos comentados en el anterior apartado 1.3. Este presente trabajo implica el desarrollo de una herramienta para la mitigación de las desigualdades socio-económicas y el cumplimiento de los derechos humanos.

Con respecto al comportamiento ético y responsabilidad social, se ha demostrado que uno de los factores influyentes en los partos prematuros es el estrato socio-económico de la madre. Siendo así que una de las variables más influyentes en los modelos es el nivel educativo de la madre, pudiendo corresponderse con el estrato social al que pertenece. No solamente con este estudio se van a tener en cuenta este tipo de variables, sino que servirá para dar más visibilidad a la necesidad de políticas de igualdad social y la garantía

del acceso a los servicios sanitarios. Además, siendo este producto una aplicación web de fácil acceso, se permite el uso de esta aplicación en la mayoría de regiones donde hay más incidencia de partos prematuros. Estas regiones suelen coincidir con países con un nivel económico bajo.

Por otro lado, con este trabajo se ha facilitado la garantía del cumplimiento de los derechos humanos. Se ha obtenido una aplicación web que permite facilitar el estudio y seguimiento de las madres con probabilidades de un parto prematuro. Así se dota al personal sanitario de una herramienta de fácil manipulación y acceso que permite una aplicación global. Además, este estudio ha puesto en el foco a la madre pudiendo así ampliar información sobre aspectos de salud que afectan a las mujeres. Se ha demostrado que es necesario seguir investigando de una manera más amplia las causas del parto prematuro, teniendo también en cuenta más aspectos de la vida de la madre (incluyendo aspectos sociales).

Por último, con respecto al apartado de sostenibilidad no se ha podido aportar ningún efecto ya que la única variable relacionada con este aspecto no influye tanto en ninguno de los conjuntos de datos. A pesar de ello, se deberían plantear nuevos estudios que tengan en cuenta los factores ambientales de la vida de la madre. Algunos de estos factores, como la contaminación del aire o las altas temperaturas, han sido objeto de estudio en diferentes artículos^{37, 38}. Es por ello, que como el cambio climático es ya una realidad; se prestan necesarios nuevos enfoques en el estudio de este tipo de variables y la afección en los partos prematuros.

Con el cumplimiento de todos los objetivos de este trabajo, se han podido garantizar los impactos previstos en este aspecto. Se cumple así las predicciones de impacto y las descripciones que se realizaron en el apartado 1.3. Por la tipología de este trabajo y teniendo ya una base de datos descrita y específica, no han aparecido ningún tipo de impacto nuevo durante el desarrollo de este mismo trabajo.

5.4. Líneas de futuro

Con este trabajo se ha podido estipular un conjunto de datos que describe las 10 variables más importantes con respecto a los partos prematuros. De esta forma, se ha facilitado la creación de una aplicación web con un cuestionario pudiendo relacionarlo con el modelo de predicción más correcto para estos datos. Aunque se haya mostrado la importancia de los factores socio-económicos y se hayan incluido algunas de este tipo de variables en el modelo final de predicción, en futuros trabajos se podría considerar la opción de ampliar estas variables. Así pues, se podría describir un perfil más amplio de características de la madre. Se podría enfocar en ampliar el número de variables a tener en cuenta usando el conjunto de datos de este mismo estudio o plantear un nuevo estudio para así crear un nuevo cuestionario con nuevos datos.

Siguiendo el hilo de esta propuesta, en este nuevo estudio se podría considerar la opción de incluir nuevas variables que estén relacionadas con la sostenibilidad y el cambio climático. Este nuevo estudio podría ampliar la información sobre cómo los factores ambientales afectan durante el embarazo pudiendo llegar a ser factores de riesgo para un parto prematuro. Se pondría en el foco la problemática actual del cambio climático y los nuevos riesgos a los que la población se enfrenta. Este estudio daría una información

amplia para la prevención desde una manera temprana de los problemas que derivan debidos a la calidad del aire, temperaturas altas, etc.

Este conjunto de datos tiene una versatilidad muy amplia debido al número y tipo de variables pudiendo extraer de estos diferentes estudios solamente centrando la atención en aspectos concretos de la madre (como modo de vida). De hecho, se han realizado algunos estudios usando este conjunto de datos seleccionando solamente algunas variables^{96,97}. Es por ello, que otra vía de estudio podría ser la exploración de más modelos de predicción centrando la atención en variables que tengan en cuenta un aspecto de la vida de la madre.

Como se puede observar, partiendo de este conjunto de datos tan amplio se pueden realizar diferentes estudios siendo así que las líneas de futuro de este trabajo son muy diversas.

6. Glosario

SVM	Support Vector Machine
kNN	K-nearest neighbor
RF	Random Forest
DT	Decision Tree
ANN	Artificial Neural Network
IDE	Integrated development environment
OMS	Organización Mundial de la Salud
NU	Naciones Unidas
UNFPA	United Nations Population Fund
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
PCA	Principal Component Analysis
MCA	Multiple Correspondence Analysis

7. Bibliografía

1. WHO. 2018. *Preterm birth*. <http://www.who.int/en/news-room/fact-sheets/detail/preterm-birth>
2. WHO, March of Dimes, PMNCH, Save the Children. 15 million preterm births: Priorities for action based on national, regional and global estimates. In: Howson CP, Kinney MV, Lawn J, eds. *Born Too Soon: The Global Action Report on Preterm Birth*. 2012.
3. United Nations Inter-agency Group for Child Mortality Estimation (UN IGME). 'Levels & Trends in Child Mortality: Report. Estimates developed by the United Nations Inter-agency Group for Child Mortality Estimation'. New York, NY: United Nations Children's Fund; 2019
4. Liu L, Oza S, Hogan D, et al. Global, regional, and national causes of under-5 mortality in 2000–15: An updated systematic analysis with implications for the Sustainable Development Goals. *Lancet*. 2016;388:3027–3035
5. Behrman RE, Butler AS, editors. *Preterm birth: causes, consequences, and prevention*. Washington DC: National Academies Press; 2007.
6. Delobel-Ayoub M, Arnaud C, White-Koning M, Casper C, Pierrat V, Garel M, et al. Behavioral problems and cognitive performance at 5 years of age after very preterm birth: the EPIPAGE Study. *Pediatrics* 2009;123:1485e92.
7. Frey HA, Klebanoff MA. The epidemiology, etiology, and costs of preterm birth. *Semin Fetal Neonatal Med*. 2016 Apr;21(2):68-73. doi: 10.1016/j.siny.2015.12.011. Epub 2016 Jan 11. PMID: 26794420.
8. Stark GF, Hart GR, Nartowt BJ, Deng J. Predicting breast cancer risk using personal health data and machine learning models. *PLoS One*. 2019 Dec 27;14(12):e0226765. doi: 10.1371/journal.pone.0226765. PMID: 31881042; PMCID: PMC6934281.
9. Dong J, Feng T, Thapa-Chhetry B, Cho BG, Shum T, Inwald DP, Newth CJL, Vaidya VU. Machine learning model for early prediction of acute kidney injury (AKI) in pediatric critical care. *Crit Care*. 2021 Aug 10;25(1):288. doi: 10.1186/s13054-021-03724-0. PMID: 34376222; PMCID: PMC8353807.
10. Universitat Oberta de Catalunya. (n.d.). *Impacte global #Agenda2030*. Uoc.edu. Retrieved June 1, 2023, from <https://www.uoc.edu/portal/ca/compromis-social/index.html>
11. *United Nations*. (n.d.). *Www.un.org*. Retrieved June 1, 2023, from <https://www.un.org/es/impacto-académico/sostenibilidad>
12. Behrman, R. E., Butler, A. S., & Institute of Medicine (US) Committee on Understanding Premature Birth and Assuring Healthy Outcomes. (2007). *The role of environmental toxicants in preterm birth*. National Academies Press.
13. Prakash, P. (2023, May 12). *Explained*. Thehindu.com. <https://www.thehindu.com/sci-tech/health/explained-india-recorded-maximum-preterm-births-in-2020-findings-of-who-report/article66838513.ece>
14. *Preterm birth complications leading cause of child mortality in India*. (2021, November 18). Thehindu.com. <https://www.thehindu.com/news/national/telangana/preterm-birth-complications-leading-cause-of-child-mortality-in-india/article37568310.ece>

15. Kundu, T. (2022, May 30). The multiple faces of inequality in India. *The Conversation*. <http://theconversation.com/the-multiple-faces-of-inequality-in-india-182074>
16. *Diccionario de cáncer del NCI*. (2011, February 2). Instituto Nacional del Cáncer. <https://www.cancer.gov/espanol/publicaciones/diccionarios/diccionario-cancer/def/diversidad>
17. España, A. I. (n.d.). *¿Qué son los derechos humanos?* Amnesty.org. Retrieved June 1, 2023, from <https://www.es.amnesty.org/en-que-estamos/temas/derechos-humanos/>
18. United Nations. (n.d.-a). *La Declaración Universal de Derechos Humanos | Naciones Unidas*. Retrieved June 1, 2023, from <https://www.un.org/es/about-us/universal-declaration-of-human-rights>
19. Salas, B. L., Urbietta, C. T., Farhane Medina, N. Z., & Castillo-Mayén, R. (2021, May 6). La salud de las mujeres y la de los hombres son distintas, pero se tratan igual. *The Conversation*. <http://theconversation.com/la-salud-de-las-mujeres-y-la-de-los-hombres-son-distintas-pero-se-tratan-igual-159950>
20. Jiang T, Gradus JL, Rosellini AJ. Supervised Machine Learning: A Brief Primer. *Behav Ther*. 2020 Sep;51(5):675-687. doi: 10.1016/j.beth.2020.05.002. Epub 2020 May 16. PMID: 32800297; PMCID: PMC7431677.
21. Biswas A, Saran I, Wilson FP. Introduction to Supervised Machine Learning. *Kidney360*. 2021 Mar 3;2(5):878-880. doi: 10.34067/KID.0000182021. PMID: 35373058; PMCID: PMC8791341.
22. (N.d.). *ieee-dataport.org*. Retrieved March 16, 2023, from <https://ieee-dataport.org/open-access/mother's-significant-feature-msf-dataset>
23. *¿Qué es el aprendizaje supervisado?* (n.d.). Ibm.com. Retrieved March 16, 2023, from <https://www.ibm.com/mx-es/topics/supervised-learning>
24. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. *BMC Med Inform Decis Mak*. 2019 Dec 21;19(1):281. doi: 10.1186/s12911-019-1004-8. PMID: 31864346; PMCID: PMC6925840.
25. Python Machine Learning - Second Edition: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow de Sebastian Raschik.
26. Blokdyk, G. (2018). *Tensorflow: A Complete Guide*. 5starcooks.
27. *Keras: Deep learning for humans*. (n.d.). Keras.io. Retrieved June 1, 2023, from <https://keras.io>
28. *DataCamp*. (n.d.). Datacamp.com. Retrieved June 1, 2023, from <https://www.datacamp.com>
29. *GeeksforGeeks*. (n.d.). GeeksforGeeks. Retrieved June 1, 2023, from <https://www.geeksforgeeks.org>
30. Pouget, ©. Unicef/raphael. (2023, May 9). 'Silent emergency': Premature births claim a million lives yearly. UN News. <https://news.un.org/en/story/2023/05/1136512>
31. *150 million babies born preterm in the last decade*. (n.d.). Unicef.org. Retrieved June 7, 2023, from <https://www.unicef.org/press-releases/150-million-babies-born-preterm-last-decade>
32. Ohuma E, Moller A-B, Bradley E (in press). National, regional, and worldwide estimates of preterm birth in 2020, with trends from 2010: a systematic analysis. *Lancet*. 2023.
33. 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>)

34. Global Trends: Forced displacement in 2021. Copenhagen: United Nations High Commissioner for Refugees; 2022.
35. Bendavid E, Boerma T, Akseer N, Langer A, Malembaka EB, Okiro EA, et al. The effects of armed conflict on the health of women and children. *Lancet*. 2021;397(10273):522-32.
36. State of global air 2020. Special Report. Boston: Health Effects Institute; 2020.
37. Chersich MF, Pham MD, Areal A, Haghighi MM, Manyuchi A, Swift CP, et al. Associations between high temperatures in pregnancy and risk of preterm birth, low birth weight, and stillbirths: systematic review and meta- analysis. *BMJ*. 2020;371.
38. McElroy S, Ilango S, Dimitrova A, Gershunov A, Benmarhnia T. Extreme heat, preterm birth, and stillbirth: A global analysis across 14 lower-middle income countries. *Environ Int*. 2022;158:106902.
39. Minckas N, Medvedev MM, Adejuyigbe EA, Brotherton H, Chellani H, Estifanos AS, et al. Preterm care during the COVID-19 pandemic: a comparative risk analysis of neonatal deaths averted by kangaroo mother care versus mortality due to SARS-CoV-2 infection. *EClinicalMedicine*. 2021;33:100733.
40. World Economic Outlook Countering the Cost of Living Crisis. Washington, D.C.: International Monetary Fund; 2022.
41. Bliss. Soaring cost of living leaves parents of sick children fearful about running medical equipment (website). London: Bliss; 2022 (<https://www.bliss.org.uk/news/2022/soaring-cost-of-living-leaves-parents-of-sick-children-fearful-about-running-medical-equipment> accessed 4 April 2023)
42. Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet* 2020;396(10258):1204-22 (<https://vizhub.healthdata.org/gbd-compare/>)
43. 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.
44. 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>)
45. Health Inequality Assessment Toolkit. World Health Organization (<https://whoequity.shinyapps.io/heat/#heat-section-2>, accessed 17 April 2023)
46. 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>)
47. 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/>)
48. Vogel JP, Ramson J, Darmstadt GL, Qureshi ZP, Chou D, Bahl R, et al. Updated WHO recommendations on antenatal corticosteroids and tocolytic therapy for improving preterm birth outcomes. *Lancet Glob Health*. 2022;10(12):e1707-e8.

49. WHO recommendation on tocolytic therapy for improving preterm birth outcomes. Geneva: World Health Organization; 2022. (<https://apps.who.int/iris/handle/10665/363128>)
50. McDonald SJ, Middleton P, Dowswell T, Morris PS. Effect of timing of umbilical cord clamping of term infants on maternal and neonatal outcomes. *Cochrane Database Syst Rev*. 2013(7):CD004074.
51. Greensides D, Robb-McCord J, Noriega A, Litch JA. Antenatal corticosteroids for women at risk of imminent preterm birth in 7 sub-Saharan African countries: a policy and implementation landscape analysis. *Global Health: Science and Practice*. 2018;6(4):644-56.
52. 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.
53. 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](https://doi.org/10.1016/s0002-9378(98)70439-9)
54. 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>
55. Burine L, Polónia D, Gradim A, editors. *How Health Data Are Managed in Mozambique*. Cham: Springer International Publishing; 2021
56. Lawn JE, Bradley E, Lawn JE, Ohuma EO, Bradley E, Suarez I, et al. Lancet series: Small Vulnerable Newborn 2. Small babies, big risks: Global estimates of prevalence and mortality for vulnerable newborns to accelerate change and improve counting. *Lancet*. 2023. (in press).
57. Roy, B. (2019, September 3). *All about missing data handling*. Towards Data Science. <https://towardsdatascience.com/all-about-missing-data-handling-b94b8b5d2184>
58. Badr, W. (2019, January 5). *6 different ways to compensate for missing values in a dataset (data imputation with examples)*. Towards Data Science. <https://towardsdatascience.com/6-different-ways-to-compensate-for-missing-values-data-imputation-with-examples-6022d9ca0779>
59. Kang, H. (2013). The prevention and handling of the missing data. *Korean Journal of Anesthesiology*, 64(5), 402–406. <https://doi.org/10.4097/kjae.2013.64.5.402>
60. Mehrotra, K. G., Mohan, C. K., & Huang, H. (2019). *Anomaly detection principles and algorithms*. Springer International Publishing.
61. Badr, W. (2019b, March 5). *5 ways to detect outliers/Anomalies that every data scientist should know (Python Code)*. Towards Data Science. <https://towardsdatascience.com/5-ways-to-detect-outliers-that-every-data-scientist-should-know-python-code-70a54335a623>
62. Bala, P. C. (2022, July 5). *How to detect outliers in machine learning – 4 methods for outlier detection*. Freecodecamp.org. <https://www.freecodecamp.org/news/how-to-detect-outliers-in-machine-learning/>
63. Goyal, C. (2021, May 19). *Outlier detection & removal*. Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2021/05/feature-engineering-how-to-detect-and-remove->

64. *Introduction to dimensionality reduction*. (2017, June 1). GeeksforGeeks. <https://www.geeksforgeeks.org/dimensionality-reduction/>
65. Brownlee J. (2020, June 30). *Introduction to Dimensionality Reduction for Machine Learning*. Machinelearningmastery.com. Retrieved June 7, 2023, from <https://machinelearningmastery.com/dimensionality-reduction-for-machine-learning/>
66. Bisong, E. (2019b). What is machine learning? In *Building Machine Learning and Deep Learning Models on Google Cloud Platform* (pp. 169–170). Apress.
67. Lantz, B. (2023). *Machine Learning with R* - (2nd ed.). Packt Publishing.
68. Saini, A. (2021, August 29). *Decision tree algorithm - A complete guide*. Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/>
69. Navlani A. (2023). *Decision Tree Classification in Python Tutorial*. Datacamp.com. Retrieved June 7, 2023, from <https://www.datacamp.com/tutorial/decision-tree-classification-python>
70. Wikipedia contributors. (2023a, April 7). *Bootstrapping (statistics)*. Wikipedia, The Free Encyclopedia. [https://en.wikipedia.org/w/index.php?title=Bootstrapping_\(statistics\)&oldid=1148726980](https://en.wikipedia.org/w/index.php?title=Bootstrapping_(statistics)&oldid=1148726980)
71. Amat R. (2020). *Random Forest python*. (n.d.). Cienciadedatos.net. Retrieved June 7, 2023, from https://www.cienciadedatos.net/documentos/py08_random_forest_python.html
72. Zhang, Z. (2019, July 31). *Support Vector Machine explained*. Towards Data Science. <https://towardsdatascience.com/support-vector-machine-explained-8bfef2f17e71>
73. *RBF SVM parameters*. (n.d.). Scikit-Learn. Retrieved June 7, 2023, from https://scikit-learn.org/stable/auto_examples/svm/plot_rbf_parameters.html
74. McNelis, P. D. (2005). What Are Neural Networks? In *Neural Networks in Finance* (pp. 13–58). Elsevier.
75. Gallo, C. (2014). Artificial Neural Networks Tutorial. In *Encyclopedia of Information Science and Technology*, Third Edition (pp. 6369–6378). IGI Global.
76. *Configuring a neural network output layer*. (2023, May 18). Enthought, Inc. <https://www.enthought.com/blog/neural-network-output-layer/>
77. Calvo, D. (2018, December 10). *Función de coste - Redes neuronales*. Diego Calvo. <https://www.diegocalvo.es/funcion-de-coste-redes-neuronales/>
78. García, J. D. V. (2020, May 26). Redes neuronales desde cero (II): algo de matemáticas. *Iartificial.net*. <https://www.iartificial.net/redes-neuronales-desde-cero-ii-algo-de-matematicas/>
79. Mishra, A. (2018, February 24). *Metrics to Evaluate your Machine Learning Algorithm*. Towards Data Science. <https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234>
80. *pandas*. (n.d.). Pydata.org. Retrieved June 7, 2023, from <https://pandas.pydata.org>
81. *Matplotlib — visualization with python*. (n.d.). Matplotlib.org. Retrieved June 7, 2023, from <https://matplotlib.org>

82. Bisong, E. (2019). NumPy. In *Building Machine Learning and Deep Learning Models on Google Cloud Platform* (pp. 91–113). Apress.
83. Varoquaux, G., Buitinck, L., Louppe, G., Grisel, O., Pedregosa, F., & Mueller, A. (2015). Scikit-learn: Machine learning without learning the machinery. *GetMobile Mobile Computing and Communications*, 19(1), 29–33. <https://doi.org/10.1145/2786984.2786995>
84. Minewiskan. (n.d.). *Training and testing data sets*. Microsoft.com. Retrieved June 7, 2023, from <https://learn.microsoft.com/en-us/analysis-services/data-mining/training-and-testing-data-sets?view=asallproducts-allversions>
85. Beretta, L., Santaniello, A. Nearest neighbor imputation algorithms: a critical evaluation. *BMC Med Inform Decis Mak* **16** (Suppl 3), 74 (2016). <https://doi.org/10.1186/s12911-016-0318-z>
86. Barnett, V., & Lewis, T. (1994). *Outliers in Statistical Data* (3rd ed.). John Wiley & Sons.
87. *Pyod 1.0.9 documentation*. (n.d.). Readthedocs.io. Retrieved June 7, 2023, from <https://pyod.readthedocs.io/en/latest/>
88. Kuhn, M., & Johnson, K. (2021). *Feature engineering and selection: A practical approach for predictive models*. Taylor & Francis.
89. Bhalla, D. (n.d.). *A complete guide to Random Forest in R*. ListenData. Retrieved June 7, 2023, from <https://www.listendata.com/2014/11/random-forest-with-r.html>
90. Prasad, A. (2020, December 23). *Tensorflow 2 for DeepLearning - Artificial Neural Networks*. Analytics Vidhya. <https://medium.com/analytics-vidhya/tensorflow-2-for-deeplearning-artificial-neural-networks-8ec72b36f493>
91. Mishra, A. (2018, February 24). *Metrics to Evaluate your Machine Learning Algorithm*. Towards Data Science. <https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234>
92. Coppola, M. (2023, May 2). *Desarrollo web: qué es, etapas y principales lenguajes*. Hubspot.es. <https://blog.hubspot.es/website/que-es-desarrollo-web>
93. *Welcome to flask — flask documentation (2.3.X)*. (n.d.). Palletsprojects.com. Retrieved June 7, 2023, from <https://flask.palletsprojects.com/en/2.3.x/>
94. Jamal, T. (2022, October 3). *How to turn your jupyter notebook into a user-friendly web app*. Freecodecamp.org. <https://www.freecodecamp.org/news/machine-learning-web-app-with-flask/>
95. *The GitHub blog*. (n.d.). The GitHub Blog; GitHub. Retrieved June 8, 2023, from <https://github.blog>
96. Deshpande, H., & Ragha, L. (2021). Mother's lifestyle feature relevance for NICU and preterm birth prediction. *ITM Web of Conferences*, 40, 03039. <https://doi.org/10.1051/itmconf/20214003039>
97. Deshpande, H. S., & Ragha, L. (2023). A hybrid random forest-based feature selection model using mutual information and F-score for preterm birth classification. *International Journal of Medical Engineering and Informatics*, 15(1), 84. <https://doi.org/10.1504/ijmei.2023.127257>