Universidad Nacional Mayor de San Marcos
Universidad del Perú. Decana de América

Dirección General de Estudios de Posgrado
Facultad de Ciencias Matemáticas
Unidad de Posgrado

Factores que determinan el embarazo adolescente a través de las técnicas del machine learning en el Perú, 2019-2020

TESIS

Para optar el Grado Académico de Magíster en Estadística Matemática

AUTOR

Bernardo CÉSPEDES PANDURO

ASESOR

Dra. Zoraida Judith HUAMÁN GUTIÉRREZ

Lima, Perú

2024



Reconocimiento - No Comercial - Compartir Igual - Sin restricciones adicionales

https://creativecommons.org/licenses/by-nc-sa/4.0/

Usted puede distribuir, remezclar, retocar, y crear a partir del documento original de modo no comercial, siempre y cuando se dé crédito al autor del documento y se licencien las nuevas creaciones bajo las mismas condiciones. No se permite aplicar términos legales o medidas tecnológicas que restrinjan legalmente a otros a hacer cualquier cosa que permita esta licencia.

Referencia bibliográfica

Céspedes, B. (2024). *Factores que determinan el embarazo adolescente a través de las técnicas del machine learning en el Perú, 2019-2020*. [Tesis de maestría, Universidad Nacional Mayor de San Marcos, Facultad de Ciencias Matemáticas, Unidad de Posgrado]. Repositorio institucional Cybertesis UNMSM.

Metadatos complementarios

Datos de autor	
Nombres y apellidos	Bernardo Céspedes Panduro
Tipo de documento de identidad	DNI
Número de documento de identidad	17615559
URL de ORCID	https://orcid.org/0000-0002-9606-1478
Datos de asesor	
Nombres y apellidos	Zoraida Judith Huamán Gutiérrez
Tipo de documento de identidad	DNI
Número de documento de identidad	09890094
URL de ORCID	https://orcid.org/0000-0002-5533-2454
Datos del jurado	
Presidente del jurado	
Nombres y apellidos	Ofelia Roque Paredes
Tipo de documento	DNI
Número de documento de identidad	06243124
Miembro del jurado 1	
Nombres y apellidos	Roberto Carlos Fiestas Flores
Tipo de documento	DNI
Número de documento de identidad	16744141
Miembro del jurado 2	
Nombres y apellidos	Moises Enrique Tarazona Cochachin
Tipo de documento	DNI
Número de documento de identidad	43491114
Datos de investigación	

Línea de investigación	A.3.2.6. Análisis de Datos y Modelamiento de Problemas de la Sociedad (Empresa, Instituciones, Población locales, regionales y nacionales).
Grupo de investigación	No aplica.
Agencia de financiamiento	Sin financiamiento.
Ubicación geográfica de la investigación	Universidad Nacional Mayor de San Marcos. Coordenadas geográficas: 12°03′30″S 77°05′00″O.
Año o rango de años en que se realizó la investigación	Julio 2021 – Noviembre 2024
URL de disciplinas OCDE	Estadísticas, Probabilidad https://purl.org/pe-repo/ocde/ford#1.01.03

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
Universidad del Perú, DECANA DE AMÉRICA

FACULTAD DE CIENCIAS MATEMÁTICAS
VICEDECANATO DE INVESTIGACIÓN Y POSGRADO
UNIDAD DE POSGRADO

ACTA DE SUSTENTACIÓN DE TESIS PARA OPTAR EL GRADO ACADÉMICO DE MAGÍSTER

Siendo las 19:00 horas del día jueves 05 de diciembre de 2024, en el Aula 102 de la Facultad de Ciencias Matemáticas de la Universidad Nacional Mayor de San Marcos, el Jurado Examinador de tesis, presidido por la Dra. Ofelia Roque Paredes e integrado por los siguientes miembros: Mg. Roberto Carlos Fiestas Flores (Jurado Informante), Mg. Moises Enrique Tarazona Cochachin (Jurado Externo) y la Dra. Zoraida Judith Huamán Gutiérrez como Miembro Asesora; se reunieron para la sustentación de la tesis titulada: «FACTORES QUE DETERMINAN EL EMBARAZO ADOLESCENTE A TRAVÉS DE LAS TÉCNICAS DEL MACHINE LEARNING EN EL PERÚ, 2019-2020», presentada por el Bachiller Bernardo Céspedes Panduro, para optar el Grado Académico de Magíster en Estadística Matemática.

Concluida la exposición, los miembros del Jurado de Tesis procedieron a formular sus preguntas que fueron absueltas por el graduando; acto seguido se procedió a la evaluación correspondiente, según tabla adjunta, habiendo obtenido el Bachiller **BERNARDO CÉSPEDES PANDURO**, el calificativo de 18 MUY BUENO

Habiendo sido aprobada la sustentación de la Tesis, el Jurado Evaluador recomienda que el Consejo de Facultad apruebe el otorgamiento del **Grado Académico de Magíster en Estadística Matemática** al Señor Bachiller **BERNARDO CÉSPEDES PANDURO**.

Siendo las 19:50 horas, se levantó la sesión, firmando para constancia la presente Acta:

Dra. Ofelia Roque Paredes
Presidente

Mg. Roberto Carlos Fiestas Flores
Miembro

Mg. Moises Enrique Tarazona ochachin
Miembro Externo

Dra. Zoraida Judith Huamán Gutiérrez
Asesora

CERTIFICADO DE SIMILITUD

Yo Zoraida Judith Huamán Gutiérrez en mi condición de asesor acreditado con la Resolución Decanal N° 001318-2021-D-FCM/UNMSM de la tesis/monografía/informe de investigación/trabajo académico, cuyo título es Factores que determinan el embarazo adolescente a través de las técnicas del Machine Learning en el Perú , 2019-2020,

presentado por el bachiller/magíster/egresado/licenciado/estudiante Bernardo Céspedes Panduro para optar el grado/título/especialidad de Grado Académico de Magíster en Estadística Matemática

CERTIFICO que se ha cumplido con lo establecido en la Directiva de Originalidad y de Similitud de Trabajos Académicos, de Investigación y Producción Intelectual. Según la revisión, análisis y evaluación mediante el software de similitud textual, el documento evaluado cuenta con el porcentaje de 6..% de similitud, nivel **PERMITIDO** para continuar con los trámites correspondientes y para su **publicación en el repositorio institucional.**

Se emite el presente certificado en cumplimiento de lo establecido en las normas vigentes, como uno de los requisitos para la obtención del grado/ título/ especialidad correspondiente.

Firma del Asesor
DNI: 09850094
Nombres y apellidos del asesor: Zoraida Judith Huamán Gutierrez

Huella digital

Dedicatoria

A mi madre y padre por todo su amor y formación de vida. A mis hermanas/os, hijas/os y a mi esposa por su amor y por su comprensión por las horas que dejamos de compartir.

Agradecimiento

A Dios por ser mi guía y por la salud que me brinda cada día.

A la Dra. Zoraida Judith Huamán Gutiérrez, asesora de la investigación, por su compromiso personal y sus consejos que me impulsaron a concluir la investigación.

A mi familia y amigos por sus consejos y palabras de aliento.

ÍNDICE GENERAL

Pág.

Dedicatoria ……... iii
Agradecimiento …… iv
Índice de contenidos ……. v
Índice de tablas ……... vii
Índice de figuras ……. viii
Resumen ……... ix
Abstract …… x
CAPÍTULO 1: INTRODUCCIÓN……. 1
1.1 Situación problemática …… 1
1.2 Formulación del Problema…… 4
1.3 Justificación ……. 4
1.3.1 Justificación Teórica ……. 4
1.3.2 Justificación Práctica …… 5
1.4 Objetivos …….. 5
1.4.1 Objetivo General …… 5
1.4.2 Objetivos Específicos …… 5
CAPÍTULO 2: MARCO TEÓRICO…….……. 7
2.1 Antecedentes del Problema ……. 7
2.1.1 Antecedentes Internacionales ……..….. 7
2.1.2 Antecedentes Nacionales …….. 9
2.2 Bases Teóricas …… 11
2.2.1 Modelos para el estudio de la fecundidad adolescente …… 11
2.2.2 Machine Learning ..…… 13
2.2.3 Máquina de vectores soporte (SVM) …… 15
2.2.4 Regresión Logística Binaria ……...……..… 19
2.2.5 Árbol de decision …...……...……..… 31
2.2.6 AdaBoost ……...…… 32
2.2.7 Gradient Boosting ……. 33
2.2.8 XGBoost ……..…. 35
2.2.9 Bagging ……. 36
2.2.10 Random Forest ……... 36
2.2.11 ExtraTrees ……..…. 38
2.2.12 Evaluación de modelos de clasificación ……..…… 38
2.2.13 Importancia de las variables con el Algoritmo Random Forest …..… 41
2.3 Definición de términos básicos ……..……. 43
2.4 Hipótesis y variables …… 46

CAPÍTULO 3: METODOLOGÍA ……. 48
3.1 Tipo y diseño de investigación …… 48
3.2 Unidad de análisis …… 48
3.3 Población de estudio ……..…….. 48
3.4 Tamaño de muestra …….. 48
3.5 Selección de muestra …… 49
3.6 Técnicas de recolección de datos ……. 49
3.7 Análisis e interpretación de la información ……. 49
CAPITULO 4: RESULTADOS Y DISCUSIÓN …….. 52
4.1 Resultados …… 52
4.2 Discusión …… 62
CONCLUSIONES ……. 65
RECOMENDACIONES …… 66
REFERENCIAS BIBLIOGRÁFICAS …….. 67
ANEXOS……. 72

ÍNDICE DE TABLAS

Pág.
Tabla 1. Operacionalización de variables …….. 46
Tabla 2. Distribución de la variable embarazo adolescente …….. 52
Tabla 3. Características sociodemográficas y de salud de las adolescentes según embarazo. Perú: 2019 al 2020 …… 54
Tabla 4. Estadísticas descriptivas de las variables cuantitativas de las adolescentes. Años: 2019 al 2020 ……. 55
Tabla 5. Tabla de contingencia de la target para los datos de entrenamiento…….. 57
Tabla 6. Tabla de contingencia de la target para los datos de evaluación ……. 57
Tabla 7. Comparación de indicadores de los modelos propuestos ……..…… 59
Tabla 8. Accuracy para los diferentes algoritmos de Machine Learning …….. 59
Tabla 9. Comparación de los algoritmos …… 61
Tabla 10. Importancia de las variables del modelo Random Forest …….. 61

ÍNDICE DE FIGURAS

Pág.
Figura 1. Determinantes de fecundidad indirectos y próximos ……..…… 12
Figura 2. Marco teórico para el análisis de la fecundidad adolescente. Di Cesare e Rodríguez-Vignoli (2006) ……. 13
Figura 3. SVM maximiza el margen entre la frontera de decisión (hiperplano de separación) y los puntos de datos de entrenamiento …….. 15
Figura 4. Geometría de un hiperplano de separación en 2D ……. 18
Figura 5. Regresión para un problema de clasificación …… 29
Figura 6. Esquema de las características de un árbol de decisión ……. 31
Figura 7. Concepto de AdaBoost para mejorar a los aprendices débiles ……... 32
Figura 8. Concepto de bagging …….. 36
Figura 9. Concepto de Random Forest ……... 37
Figura 10. Matriz de confusión …….. 39
Figura 11. Característica operativa del receptor …… 41
Figura 12. Correlación de las variables del estudio …… 56
Figura 13. Matrices de confusión para todos los algoritmos …… 58
Figura 14. Curvas ROC para los algoritmos utilizados …… 60

RESUMEN

FACTORES QUE DETERMINAN EL EMBARAZO ADOLESCENTE A TRAVÉS DE LAS TÉCNICAS DEL MACHINE LEARNING EN EL PERÚ, 2019-2020

Bernardo Cespedes Panduro

Mayo - 2024

El objetivo del presente estudio es determinar los factores más importantes que predicen el embarazo adolescente a través de la mejor técnica del machine learning en el Perú, durante los años 2019 al 2020. Para la elección de los factores asociados a la fecundidad adolescente para su posterior análisis se hace uso del marco teórico propuesto por Di Cesare e Rodríguez-Vignoli en el año 2006.

Se propone nueve técnicas de machine learning como son: Máquina de Vectores Soporte, Regresión Logística Binaria, Árbol de decisión, AdaBoost, Gradient Boosting, XGBoost, ExtraTrees, Bagging y Random Forest para el estudio del embarazo adolescente en el Perú, eligiéndose la mejor técnica a partir de sus métricas (Accuracy y ROC AUC).

El presente trabajo de corte transversal utilizó la base de datos de la Encuesta Demográfica y de Salud Familiar (ENDES) 2019-2020 que fue dirigida por el Instituto Nacional de Estadística e Informática (INEI), utilizándose una muestra de 16825 mujeres adolescentes encuestadas entre los años 2019 al 2020.

El estudio encuentra que la mejor técnica para predecir el embarazo adolescente a partir de los factores que se han tenido en cuenta es Random Forest (Accuracy = 96.58% y AUC = 99%), por otro lado, los factores más importantes fueron: Ha tenido una primera unión, seguido por Alguna vez usó cualquier método anticonceptivo y Actualmente asiste a la escuela, colegio, instituto o universidad.

Palabras claves: Técnicas de Machine Learning, embarazo adolescente, ENDES y Accuracy.

ABSTRACT

FACTORS THAT DETERMINE ADOLESCENT PREGNANCY THROUGH MACHINE LEARNING TECHNIQUES IN PERU, 2019-2020

Bernardo Cespedes Panduro

May - 2024

The objective of this study is to determine the most important factors that predict teenage pregnancy through the best machine learning technique in Peru, during the years 2019 to 2020. In order to choose the factors associated with adolescent fertility for subsequent analysis, the theoretical framework proposed by Di Cesare and Rodriguez-Vignoli in 2006 is used.

Nine machine learning techniques are proposed such as: Support Vector Machine, Binary Logistic Regression, Decision Tree, AdaBoost, Gradient Boosting, XGBoost, ExtraTrees, Bagging and Random Forest for the study of teenage pregnancy in Peru, choosing the best technique based on its metrics (Accuracy and ROC AUC).

This cross-sectional study used the database of the Demographic and Family Health Survey (ENDES) 2019-2020 conducted by the National Institute of Statistics and Informatics (INEI), using a sample of 16825 adolescent women surveyed between 2019 and 2020.

The study finds that the best technique to predict teenage pregnancy from the factors taken into account is Random Forest (Accuracy = 96.58% and AUC = 99%), on the other hand, the most important factors were: Has had a first union, followed by Ever used any contraceptive method and Currently attending school, college, high school or university.

Keywords: Machine Learning techniques, teenage pregnancy, ENDES and Accuracy.

CAPITULO 1

INTRODUCCIÓN

1.1. Situación problemática

La tasa de fecundidad adolescente en América Latina ha mostrado un descenso gradual en comparación con la media mundial a lo largo del tiempo (Pinzón-Rondón et al., 2018). En Perú, por ejemplo, la proporción de la fecundidad adolescente con respecto a la fecundidad total ha aumentado como resultado de descensos menos significativos y oscilantes entre las mujeres de 12 a 17 años, a pesar de un descenso estable de las tasas globales de fecundidad (INEI, 2021).

La ENDES reporta que la fecundidad adolescente ha disminuido en 0,5 puntos porcentuales (de 2,8% a 2,3%) entre 2018 y 2020 (INEI, 2021). Esta disminución no es particularmente sustancial, considerando las graves consecuencias asociadas al embarazo adolescente. A nivel social, contribuye a elevar los indicadores de abandono escolar en la educación secundaria, aumenta la susceptibilidad a la participación en redes delictivas o de explotación sexual comercial (Loredo-Abdalá et al., 2017) y, en última instancia, disminuye las perspectivas de empleo a largo plazo, mermando así los ingresos económicos de la unidad familiar y fomentando un nivel de pobreza más profundo (Loredo-Abdalá et al., 2017; Williamson et al, 2013). Este fenómeno también genera importantes costes sociales (Williamson et al., 2013). Del mismo modo, en determinadas regiones y grupos sociales se observa el embarazo precoz como un fenómeno cultural; sin embargo, en las grandes ciudades, estos embarazos no suelen ser bien recibidos y solo se dan en noviazgos que no viven juntos; o también se producen en el seno de parejas consentidas que suelen culminar con la desprotección de la mujer y el niño, contribuyendo así a la cuestión social de la maternidad en solitario. Numerosos embarazos culminan en abortos practicados por personas sin formación profesional adecuada y en condiciones insalubres, debido a la escasez y alto coste de los servicios médicos especializados y al

hecho de que, según el informe ENDES 2020, el aborto está prohibido en Perú (INEI, 2021).

El embarazo durante la adolescencia disminuye la probabilidad de establecer una relación sentimental estable y aumenta la posibilidad de obtener trastornos emocionales, incluida la depresión, que en ocasiones va acompañada de ideación suicida (Williamson et al., 2013) y abuso de sustancias (Loredo-Abdalá et al., 2017). Con respecto a las ramificaciones biológicas, las adolescentes se enfrentan a una mayor susceptibilidad a diversas complicaciones en comparación con las mujeres adultas. Entre ellas se incluyen la preeclampsia (Jeha et al., 2015; Azevedo et al., 2015), la anemia (Jeha et al., 2015), el parto prematuro (Azevedo et al., 2015; Kaplanoglu et al., 2015; Williamson et al., 2013), el bajo peso al nacer (Kaplanoglu et al., 2015; Azevedo et al., 2015) y la mortalidad materna y perinatal (Azevedo et al., 2015; Williamson et al., 2013).

Los determinantes de la fecundidad son los mecanismos que explican la variación específica de la población en los niveles y la composición de la fecundidad. Estas variables influyen en la probabilidad de alcanzar la fecundidad y pueden clasificarse como directas o indirectas, que varían en función del carácter de su impacto en la fecundidad. El desarrollo científico y teórico de cómo las normas sociales influyen en las prácticas tradicionales ha sido fundamental para estas consideraciones (Arriaza, 2023).

En general, la aplicación de las pruebas de la investigación parece contribuir significativamente al avance de los conocimientos y las estrategias en materia de fecundidad adolescente. Sin embargo, además de perfeccionar nuestras estrategias políticas y de intervención a la luz de estas pruebas, también debemos contemplar la mejora de nuestras metodologías de investigación para facilitar el desarrollo de hipótesis novedosas en este campo. En la actualidad, nuestros enfoques en este ámbito dependen predominantemente de análisis epidemiológicos y demográficos convencionales de datos de encuestas exhaustivas y representativas a escala nacional procedentes de naciones afectadas por la fecundidad adolescente. En dichos análisis, seleccionamos de antemano los posibles correlatos y evaluamos las hipótesis relativas a los correlatos previstos. Con ese fin, los enfoques del machine learning ofrecen una capacidad prometedora para ofrecer un enfoque de generación de hipótesis, de manera que se puedan ampliar las metodologías de investigación en salud pública, en particular para la exploración de variables novedosas (Mooney & Pejaver, 2018; Bellinger et al., 2017).

El Machine Learning, que es un campo amplio y en rápido crecimiento, es, en esencia, la aplicación de algoritmos para analizar datos y adquirir conocimientos a partir de ellos, con la intención de descubrir asociaciones entre variables, ya sea para predecir, clasificar o simplemente identificar patrones en los datos. [Nota: el término «predecir» se emplea en el machine learning para evaluar correlaciones; sin embargo, debido a la naturaleza transversal de nuestros datos, los supuestos causales implícitos en el término «predecir» no son aplicables]. En el contexto de este estudio, el término «predecir» sirve como prueba de asociación. El algoritmo de machine learning aprende el modelo óptimo a través de una serie de iteraciones, que dependen del tipo de tarea y del algoritmo seleccionados (LeCun, Bengio & Hinton, 2015; Bellinger et al., 2017; Mooney & Pejaver, 2018). Una de las modalidades clave en las que el machine learning puede diferir de los métodos epidemiológicos tradicionales es su orientación hacia la generación de conjeturas en lugar de la contrastación de hipótesis (Bellinger et al., 2017).

La utilización de métodos de machine learning para analizar conjuntos de datos masivos permite explorar numerosas variables presentes en amplios conjuntos de datos de encuestas que pueden no estar asociadas hipotéticamente con el resultado de interés especificado. Esta correlación podría atribuirse a factores espurios o a la necesidad de seguir desarrollando los fundamentos teóricos que sustentan nuestras hipótesis. Como alternativa, podemos reducir la probabilidad de identificar tales relaciones caóticas empleando algoritmos diseñados para grandes conjuntos de datos que contienen un número potencialmente elevado de variables que no están relacionadas con el resultado deseado. Los algoritmos de machine learning pueden, por lo tanto, respaldar la generación de hipótesis (Mooney & Pejaver, 2018), de manera muy similar a como lo haría la investigación cualitativa (Glaser & Strauss, 2017), lo que requiere tanto dominio como experiencia metodológica para interpretar los hallazgos para la generación de hipótesis y la expansión de la teoría. Aunque aún está en sus inicios en lo que respecta a las aplicaciones de salud pública, este concepto ya se ha puesto en práctica en diversos ámbitos, como la salud ambiental, la salud física, la salud cognitiva y los determinantes sociales de la salud (Seligman et al., 2018; DeGregory et al., 2018; Bratic et al., 2018; Bellinger et al., 2017; Daoud et al., 2019). No se ha localizado ninguna prueba empírica que apoye su aplicabilidad al tema investigado, a saber, la fecundidad adolescente.

En este estudio, se ha utilizado el machine learning para identificar variables relacionadas con el embarazo adolescente en función de los datos recogidos con la

ENDES aplicada a una muestra muy grande representativa a nivel nacional de mujeres de con una edad de 12 a 19 años en el Perú. Nos centramos en estos datos debido a los amplios análisis sobre el tema del embarazo adolescente realizados con estos datos de vigilancia en años de estudio anteriores. Una extensa investigación previa sobre el tema nos permite una sólida comprensión de los datos de la encuesta y el dominio (Raj et al., 2009). Además, este conjunto de datos posee características “amplio” (comprende un número considerable de variables) y “alto” (comprende un número considerable de sujetos), lo que permite que el aprendizaje automático resulte ventajoso para optimizar la utilización de todas las variables viables y descubrir relaciones entre variables no consideradas previamente. Por lo tanto, la exploración de este tema con este conjunto de datos de Perú permite la identificación de factores de riesgo no considerados e inexplorados a los que podemos apuntar para acelerar el cambio en caso de que un "efecto techo" arraigue en la reducción observada de la práctica.

El análisis de estos datos no solo proporcionará información sobre las variables o constructos relacionados con el embarazo adolescente que hasta ahora no se comprenden bien o no son objeto de los esfuerzos actuales, este análisis también proporciona información sobre la validez del machine learning para identificar fuertes relaciones del embarazo adolescente, dado la extensa investigación previa sobre este tema utilizando análisis tradicionales.

1.2 Formulación del Problema

¿Cuáles son los factores más importantes que predicen el embarazo adolescente a través de la mejor técnica del machine learning en el Perú, durante los años 2019 al 2020?

1.3 Justificación

1.3.1 Justificación teórica

Un indicador de la salud reproductiva que se asocia al desarrollo social de una nación es el embarazo en la adolescencia, dado que suele ir acompañado de una mayor vulnerabilidad y riesgos psicológicos, físicos y económicos (Yakubu & Salisu, 2018).

En Perú, el embarazo adolescente ha disminuido de una manera no significativa, producto posiblemente al poco interés que le presta el estado mediante su gestión pública. Asimismo, de la bibliografía revisada, no se ha encontrado estudios para Perú, que empleen técnicas de machine learning que discriminen los factores asociados al embarazo

adolescente. En ese sentido, este estudio es importante para conocer los factores que intervienen en el embarazo adolescente e implementar políticas públicas de salud.

Asimismo, este estudio mejora la comprensión de la relación entre el embarazo en menores y una multitud de factores concomitantes. Los resultados se pueden utilizar como un insumo para fundamentar las decisiones relativas a las políticas educativas y de salud pública.

Además, los responsables de las políticas de salud podrán influir en la reducción del embarazo adolescente identificando los factores asociados al mismo. Esto ayudará al Perú a alcanzar los Objetivos de Desarrollo del Milenio establecidos por las Naciones Unidas en el año 2000, que abarcan objetivos relacionados con la educación, la pobreza y la salud reproductiva, incluida la reducción de los embarazos de adolescentes y jóvenes.

Finalmente, al utilizar técnicas de machine learning se están empleando procedimientos matemáticos útiles para trabajar con grandes bases de datos, donde se prueba los resultados en una base de datos diferente a la que se empleó para entrenar el modelo obteniendo diferentes métricas a partir de la matriz de clasificación o confusión.

1.3.2 Justificación práctica

Con la ejecución de la presente tesis se hizo un recorrido por las técnicas más utilizables en el machine learning de clasificación sobre la predicción que existen permitiendo elegir la que discrimina mejor el embarazo adolescente. Con el conocimiento de la técnica que resultó mejor para discriminar el embarazo adolescente en las mujeres sirve en las decisiones administrativas relacionadas con la salud que se toman en los sectores público y privado, para de esta manera prevenir y/o gestionar recursos con la finalidad de mitigar la prevalencia de esta dolencia como es el embarazo adolescente, que genera repercusiones negativas para la sociedad en general.

La presente investigación es original, puesto que las técnicas de machine learning aplicadas al embarazo adolescente y como prevenirlo es nuevo en nuestro país.

1.4 Objetivos

1.4.1 Objetivo General

Determinar los factores más importantes que predicen el embarazo adolescente a través de la mejor técnica del machine learning en el Perú, durante los años 2019 al 2020.

1.4.2 Objetivos Específicos

- Describir los factores intermedios y subyacentes según embarazo adolescente.
- Calcular y comparar las métricas que permiten predecir el embarazo adolescente en el Perú a partir de las matrices de confusión obtenidas con las técnicas de machine learning.
- Establecer la mejor técnica de machine learning que determina los factores que inciden en el embarazo adolescente del Perú, durante los años 2019 al 2020.

CAPITULO 2

MARCO TEÓRICO

2.1 Antecedentes del Problema

2.1.1 Antecedentes Internacionales

En su publicación, Quezada et al. (2020) utilizaron técnicas de minería de datos para establecer los determinantes socioeconómicos más significativos del embarazo adolescente en Colombia. El estudio utilizó una muestra de 700 adolescentes, siendo la variable dependiente si tenían o no hijos. También se analizaron 11 variables independientes, empleando el algoritmo J48 de la plataforma Weka. Los resultados indican que las variables socioeconómicas más influyentes en el embarazo adolescente son los siguientes: pareja actual, aval de métodos anticonceptivos por parte de la entidad promotora de salud (EPS), escolaridad actual, inicio de actividad sexual a partir de los quince años, planes de actividad sexual, nivel educativo (con una eficacia superior al 97%) y tipo de educación.

El objetivo del artículo de Rosales (2019) es examinar los factores indirectos e intermedios que influyen en la fecundidad adolescente en Honduras y su conexión con el nivel de educación. Se reproduce metodológicamente el modelo de fecundidad planteado por John Bongaarts, empleando datos obtenidos de la Encuesta Demográfica y de Salud (ENDESA) realizada en 2011-2012. Tras el análisis de las variables intermedias propuestas en el modelo, la nupcialidad o unión precoz aparece como el factor más influyente; la tasa global de fecundidad adolescente en un periodo de tres años es de 1,46, lo que representa el número medio de hijos por cada mil mujeres adolescentes. Además, la encuesta reveló que la mayoría de las adolescentes embarazadas y las que ya habían sido madres en el momento del estudio residían en regiones rurales. Esas adolescentes se diferencian de la población urbana por atributos sociodemográficos, como mayor población infantil, menor nivel de educación y residencia en los índices de riqueza más inferiores.

En su artículo, Albornoz-Arias et al. (2019) presentan un análisis que utiliza un enfoque socioecológico para examinar la relación entre diversas variables asociadas al embarazo adolescente: (a) conocimiento del período fértil en relación con el nivel educativo, lugar de residencia y pareja conviviente de la adolescente, (b) decisión de embarazarse, (c) uso del condón al momento del inicio de las relaciones sexuales, (d) circunstancias que rodean el embarazo no deseado. La muestra comprendió 1963 madres adolescentes, las que fueron seleccionadas empleando el muestreo aleatorio estratificado de asignación proporcional sobre una población de mujeres adolescentes residentes en el estado Táchira. Se aplicó una encuesta estructurada a una muestra intencional, no probabilística e independiente, mientras que un cuestionario titulado "Maternidad adolescente en el Estado de Táchira" fue completado por 235 adolescentes embarazadas en su cuarto mes o subsiguiente. Los resultados del estudio indican que la elección de la pareja no influye en la decisión de la adolescente de quedarse embarazada; no hubo indicios de correlación entre la probabilidad de quedarse embarazada y el nivel de escolaridad ni entre la edad y el uso del preservativo.

El objetivo del artículo de Albornoz-Arias et al. (2018) es delinear la correlación entre diversas variables determinantes distales que se asocian con el embarazo adolescente: (a) Consentimiento para quedarse embarazada, (b) Utilización del preservativo durante la primera relación sexual, y (c) Conocimiento del periodo fértil teniendo en cuenta el nivel de escolaridad de la adolescente, la situación de convivencia con su pareja y el lugar de residencia; (d) Circunstancias que rodean al embarazo no deseado; (e) Embarazo no deseado; (f) Mantener relaciones sexuales sin anticonceptivos con la creencia de que no ocurriría nada; (g) Acto sexual intentado con anticonceptivos actuales pero sin éxito, teniendo en cuenta la situación de pareja de la adolescente y su nivel de educación. La población del estudio comprendió 6435 mujeres de 10 a 19 años, incluyendo adolescentes embarazadas por lo menos durante el cuarto mes y madres con hijos de a lo mucho 18 meses de edad. Los resultados pertinentes indican una correlación moderada entre el grado de estudios y el estado conyugal; una asociación significativa y robusta entre el estado civil y el lugar de residencia de la adolescente, dependiendo de si ella y su pareja comparten o no residencia; y una asociación entre las circunstancias que llevaron al embarazo no deseado y la situación conyugal.

Pinzón-Rondón et al. (2018) son autores de un artículo en el que delinean los factores familiares, individuales y de comportamiento sexual que se correlacionan con el

embarazo adolescente y el inicio de la actividad sexual en mujeres colombianas de 13 a 19 años en 2010. Se empleó un estudio transversal, que incorpora un componente analítico, para recolectar datos de una muestra de mujeres 13313 de 13 a 19 años que participaron en la ENDES 2010. Los investigadores utilizaron regresión logística multinomial para su análisis. Un examen mediante análisis multivariante reveló que la edad, la cohabitación, la exposición a la violencia sexual y la violencia de los padres eran factores de riesgo significativos para el inicio de las relaciones sexuales y también para el embarazo en la adolescencia. Se determinó que la asistencia a la escuela proporcionaba protección contra ambos incidentes. La pobreza, la superpoblación y la falta de educación, aunque no estaban relacionadas con el inicio de la actividad sexual, sí lo estaban con el embarazo en la adolescencia.

En la publicación de su artículo, Binstock (2017) examina el impacto de los modelos de historia de eventos discretos sobre la probabilidad de maternidad adolescente a través del lente de la educación, la edad y circunstancias de inicio sexual y uso de anticonceptivos, y el inicio de relaciones protoafectivas. El artículo se deriva de datos obtenidos de una encuesta del año 2012 que abarcó a 1571 mujeres residentes en las regiones sanitarias V y VII de la provincia de Buenos Aires, además de las ciudades capitales de Chaco y Santiago del Estero (480 adolescentes de 18-19 años y 1091 mujeres jóvenes de 20-24 años de edad). Sus conclusiones indican que el contexto de crianza, la edad a la que se produce el inicio sexual y si se toman precauciones para prevenir el embarazo durante el inicio sexual son predictores significativos de la maternidad adolescente. Del mismo modo, se demuestra la importancia de la continuidad educativa en el retraso de la maternidad y el funcionamiento de la institución educativa en la transmisión de conocimientos, especialmente durante los años de formación que preceden a la iniciación sexual.

2.1.2 Antecedentes Nacionales

El objetivo del artículo de Castañeda y Santa-Cruz-Espinoza (2021) era evaluar los determinantes de riesgo asociados al embarazo en la adolescencia. En este estudio retrospectivo de casos y controles se utilizó una muestra de conveniencia no probabilística de 180 adolescentes; siendo los casos: 60 embarazadas y los controles: 120 no embarazadas. Utilizando un formulario de identificación de datos sociodemográficos y familiares y la Escala de Evaluación de la Cohesión y Adaptabilidad Familiar (FACES-III), se recogieron datos en un centro de salud pública de Trujillo. Los investigadores

descubrieron que los factores de riesgo moderadamente significativos incluían el funcionamiento de la familia, la violencia, el estado sentimental del padre y la edad de la madre al primer embarazo. Por el contrario, el nivel educativo de la madre tenía un efecto de pequeña magnitud, y el nivel educativo alcanzado por el padre no era estadísticamente significativo.

D'Añari (2019) realizó un artículo con el objetivo de establecer una asociación entre los determinantes biosociodemográficos y el embarazo en adolescentes gestantes que fueron atendidas por el Servicio de Obstetricia del Hospital Regional Honorio Delgado Espinoza entre abril y mayo del 2019. Se encuestó a todas las gestantes adolescentes que cumplieron con los criterios de inclusión y fueron atendidas en el servicio en mención. El estudio utilizó la prueba no paramétrica chi cuadrado para conocer las relaciones entre las variables investigadas. Según los resultados, se observó una asociación estadística entre el embarazo precoz y las siguientes variables: nivel de estudios, consumo de alcohol, número de parejas sexuales, edad de inicio de las relaciones sexuales y organización familiar.

En su tesis, Ávalos (2018) planteó como objetivo identificar los factores que se asocian al embarazo precoz que fueron atendidas en el Centro de Salud Bellavista I-4 Sullana en el año 2017, siendo su población todas las adolescentes gestantes que fueron atendidas en este Centro. La muestra fue de 55 adolescentes embarazadas y se empleó la prueba chi cuadrado para encontrar los factores que se asocian al embarazo precoz, siendo ellos: la menarquia después de los 13 años y las relaciones sexuales iniciadas antes de los 16 años.

En su artículo del año 2016, Onofre y Vargas examinaron los determinantes que influyen en la fecundidad adolescente de 15 a 19 años de edad en el hospital "Carlos Monge Medrano" de Juliaca. Se empleó una ficha de datos económicos, sociales y demográficos para recolectar información de una muestra de 230 adolescentes que ingresaron al hospital. Se encontró que los determinantes estaban asociados con la fertilidad mediante la utilización de la prueba de chi cuadrado. Se descubrió que el 0,6652 de las adolescentes eran fértiles. Proporciones comparables de adolescentes ya eran madres en todos los grupos de edad: 0,6250 de las adolescentes de 15 años y 0,6667 de las adolescentes de 19 años. El 0,6950 de las adolescentes que cohabitaban eran madres, frente al 0,4667 de las adolescentes solas; se trata de una diferencia estadísticamente significativa. La fecundidad fue del 0,7934 entre las adolescentes que no residían con sus

padres, del 0,7647 entre las que vivían con un solo progenitor (el padre o la madre) y del 0,4783 entre las adolescentes que vivían con sus padres. Las tasas de fecundidad de las distintas categorías educativas eran comparables, excepto en el caso de las adolescentes con estudios superiores, cuya tasa de fecundidad era del 0,9231, una divergencia sustancial con respecto a los demás grupos. La fecundidad no difería según el tipo de acceso a los servicios de salud reproductiva; pero, las adolescentes con mejor acceso tenían una tasa de concepción más alta que las que tenían un acceso inadecuado. La fecundidad fue mayor entre las adolescentes de familias con ingresos más altos (71,01%) que entre las de familias con ingresos más bajos (64,29%). Las tasas de fecundidad de las adolescentes variaban según la región: 57,35% en áreas rurales, 82,95% en áreas periurbanas y 55,41% en áreas urbanas. Llegaron a la conclusión de que el estado civil, la organización familiar y el área de residencia actúan en la elevada fecundidad adolescente.

2.2 Bases Teóricas

2.2.1 Modelos para el estudio de la fecundidad adolescente

La investigación inicial sobre los determinantes que actúan en la fecundidad corrió a cargo de Davis & Blake (1956), que introdujeron un modelo compuesto por once variables intermedias clasificadas en tres grupos y que influían en la fecundidad (Davis & Blake, 1956). A continuación, Bongaarts (1978) identificó siete factores que podían influir directamente en la probabilidad de que se produjera la fecundidad; los denominó determinantes intermedios de la fecundidad. El autor identificó cuatro de estos factores como cruciales: el porcentaje de mujeres casadas; la utilización y eficacia de los anticonceptivos; la duración del posparto; y el aborto inducido. Dado que este modelo fue desarrollado inicialmente para las mujeres en edad de reproducción y se está aplicando también al estudio de la fecundidad adolescente, es necesario modificarlo para ajustarlo a las características de esta población (Rodríguez Vignoli, 2014).

Di Cesare y Rodríguez-Vignoli (2006) introdujeron un modelo que intenta dilucidar los determinantes de la fecundidad en la población adolescente, concretamente en la de 15 a 19 años. Los autores plantean un modelo que, inspirándose en el de Bongaarts (1978), categoriza los factores que influyen en la fecundidad en dos grupos: a) variables intermedias (determinantes próximos), que pertenecen a los factores más cercanos al sujeto; y b) factores subyacentes (determinantes indirectos), que están

asociados a variables socioculturales, ambientales, familiares e individuales (ver Figura 1). Este modelo incorpora las variables intermedias posteriores: edad de la primera relación sexual, uso de anticonceptivos y estado civil. Los determinantes que influyen en la variación de los niveles de fecundidad a través de determinantes próximos de la fecundidad adolescente se clasifican en tres categorías: socioculturales, familiares e individuales (Di Cesare y Rodríguez Vignoli, 2006).



Figura 1. Determinantes de fecundidad indirectos y próximos.
Fuente: Bongaarts (1978).

La figura 2 ilustra el esquema propuesto por Di Cesare y Rodríguez-Vignoli (2006) en relación con las variables intermedias y los factores subyacentes de la fecundidad adolescente. Esta figura ilustra la interacción entre las variables intermedias y los factores subyacentes que actúan en la probabilidad de la fecundidad adolescente. Dado que los determinantes subyacentes no ejercen una influencia directa sobre la probabilidad de fecundidad adolescente, su actuación depende de la influencia de las variables intermedias. A la inversa, las variables intermedias tienen la capacidad de ejercer una influencia total sobre la variabilidad de la fecundidad adolescente; sin embargo, su papel típico es servir de intermediarias entre los efectos de los factores subyacentes. Por lo tanto, es habitual que la probabilidad de la fecundidad adolescente se vea influida por una interacción entre las variables intermedias y los factores subyacentes (Di Cesare y Rodríguez Vignoli, 2006).

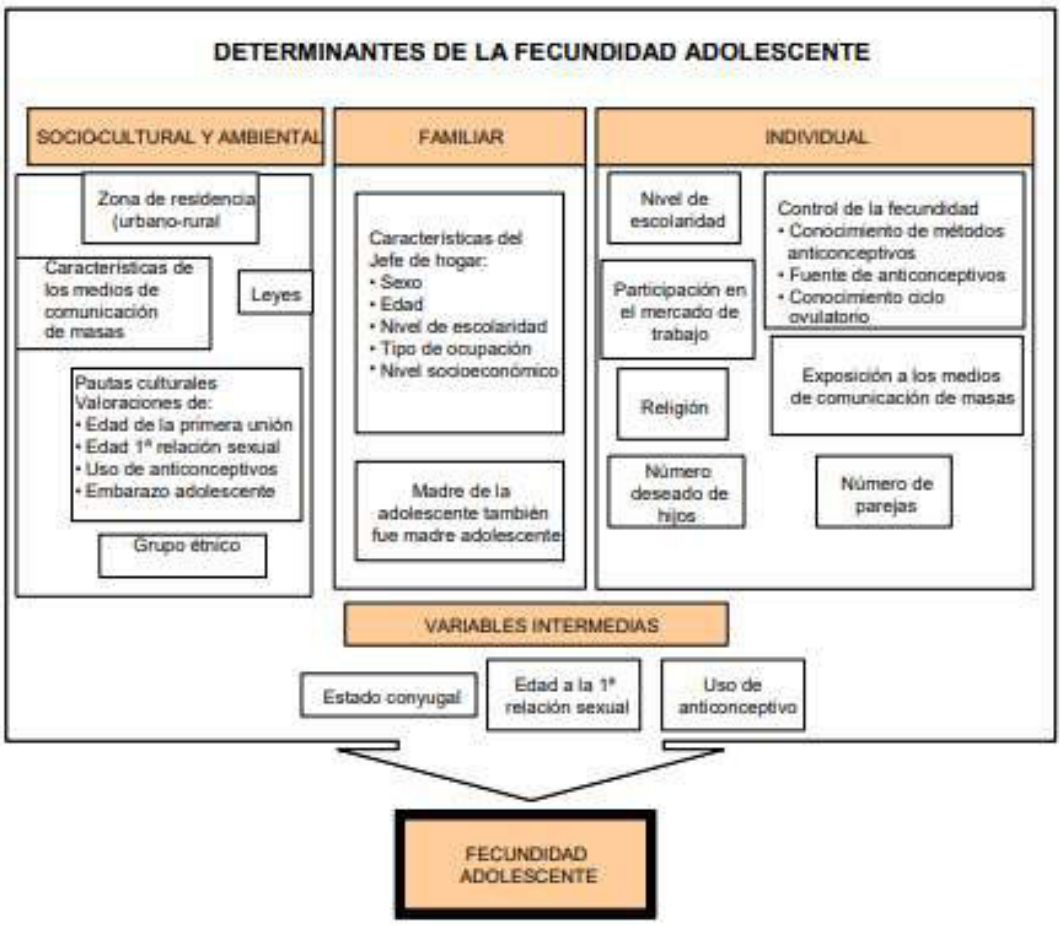


Figura 2. Marco teórico para el análisis de la fecundidad adolescente. Di Cesare e Rodríguez-Vignoli (2006)

La presente investigación utilizó el modelo de Di Cesare y Rodríguez-Vignoli (2006), que es el más reciente para el estudio de la fecundidad adolescente y está basado en el modelo de variables intermedias y factores subyacentes de Bongaarts (1978) cuyo orden de acción sobre la fecundidad es diferente.

2.2.2. Machine Learning

En 1959, Arthur Samuel, uno de los primeros defensores de la inteligencia artificial, dio la definición inicial de Machine Learning. Como sugiere esta definición, el Machine Learning es un dominio dentro de la inteligencia artificial que pretende dotar a las máquinas que sean capaces de aprender y entrenarse a partir de datos sin programación explícita. Emula la forma en que los humanos adquieren conocimientos observando ejemplos, sin basarse en fórmulas o reglas entre variables, y en última instancia desarrolla un modelo capaz de generalizar -es decir, producir resultados en situaciones novedosas- mediante el uso de algoritmos. Este comportamiento es crítico porque permite resolver

problemas en los que localizar una fórmula que dé una respuesta exacta basada en el conocimiento de variables específicas es imposible o extremadamente difícil. Mediante este enfoque, el modelo puede ser "entrenado" para identificar y categorizar imágenes novedosas; esto ocurre cuando es expuesto a una serie de casos o ejemplos que corresponden a atributos específicos y a la imagen correspondiente. Así, el Machine Learning surge como una ciencia algorítmica que permite a las máquinas adquirir la capacidad de resolver problemas que son inviables de resolver manualmente.

De manera fundamental, se puede clasificar el machine learning en dos tipos:

- ***El aprendizaje supervisado***. El objetivo de esta forma de aprendizaje es proporcionar una explicación para una o más variables dependientes utilizando otras variables independientes. Las entradas o valores de las variables independientes y las salidas o valores de las variables dependientes informan los algoritmos asociados a este tipo de aprendizaje. Formados por observaciones observables, los inputs y outputs funcionan como "supervisores" de los resultados adquiridos durante el proceso de aprendizaje.

  Los modelos que corresponden a este tipo de aprendizaje y ayudan a la predicción se denominan *modelos predictivos*.
- ***El aprendizaje no supervisado***. Se denomina aprendizaje no supervisado a la búsqueda de relaciones entre un conjunto de variables en ausencia de una variable dependiente. Este enfoque de aprendizaje genera modelos descriptivos, a diferencia del aprendizaje supervisado, que genera modelos predictivos.

Cuando Y es una variable categórica en el contexto de los modelos supervisados, el modelo se denomina “modelo de *clasificación*”. La denominación de un modelo en el que Y es una variable continua es "modelo de *regresión*".

Considerando que la variable dependiente es de tipo cualitativa nominal, este estudio emplea modelos de clasificación supervisada. El objetivo es construir un modelo que asigne cada elemento de la población a una de las categorías “Y”, de acuerdo con las categorías establecidas por las variables independientes, que pueden ser de naturaleza numérica o categórica.

En el presente estudio se utilizó los modelos supervisados de clasificación: Máquina de Soporte Vectorial, Regresión Logística Binaria, Árbol de decisión, AdaBoost, Gradient Boosting, XGBoost, ExtraTrees, Random Forest y Bagging.

2.2.3. Máquina de vectores soporte (SVM)

Son un tipo de algoritmo de clasificación que utiliza discriminantes lineales de margen máximo. En otras palabras, el objetivo es identificar el hiperplano que maximiza eficazmente el margen o espacio entre clases. Además, podemos utilizar el truco del kernel para encontrar el límite de decisión no lineal óptimo entre las clases, que corresponde a un hiperplano en un espacio "no lineal" de alta dimensión.

Los "vectores de soporte" asignados son los ejemplos de entrenamiento más cercanos al hiperplano de separación (límite de decisión), a partir del cual el margen se calcula como la distancia.

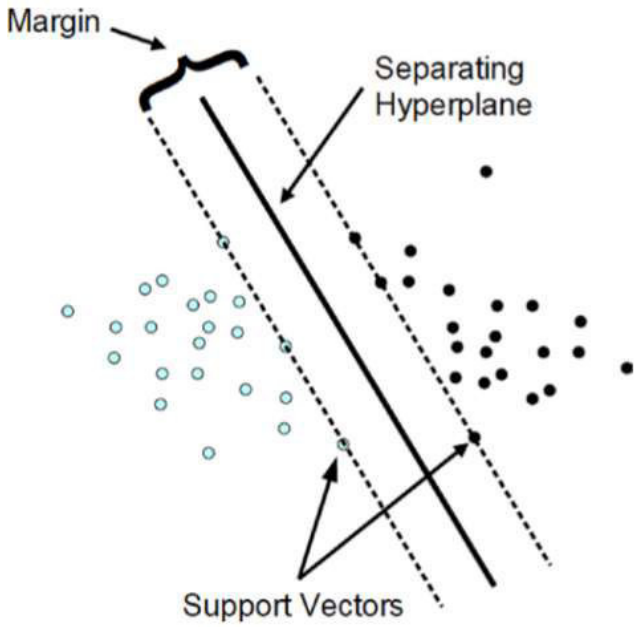


Figura 3. SVM maximiza el margen entre la frontera de decisión (hiperplano de separación) y los puntos de datos de entrenamiento.

Unos márgenes amplios en las fronteras de decisión indican un error de generalización reducido, mientras que los modelos con márgenes pequeños son más susceptibles de sobreajuste (Raschka et al., 2023).

Vectores de soporte y márgenes

Sea D un conjunto de datos de clasificación, con *n* puntos x_i en un espacio d-dimensional. Además, supongamos que sólo hay dos etiquetas de clase, es decir, $y_i \in \{+1,-1\}$, que denotan las clases positiva y negativa.

Hiperplanos

Un hiperplano en *d* dimensiones viene dado como el grupo de todos los puntos x $\in \mathbb{R}^d$ que satisfacen la ecuación $h(x) = 0$, donde $h(x)$ es la función hiperplano, definida como sigue:

$$h(x) = w^T x + b = w_1 x_1 + w_2 x_2 + \cdots + w_d x_d + b$$

Aquí, w es un *vector de pesos d* dimensional y b es un escalar, llamado *sesgo*. Para los puntos que se encuentran en el hiperplano, tenemos

$$h(x) = w^T x + b = 0$$

El hiperplano se determina así como el grupo de todos los puntos tales que $w^T x = -b$.

Recordemos además que el vector de pesos w es ortogonal o normal al hiperplano, mientras que $-\frac{b}{w_i}$ especifica el desplazamiento en el que el hiperplano interseca la i-ésima dimensión (siempre que $w_i \neq 0$). En otras palabras, el vector de peso w especifica la dirección que es normal al hiperplano, que fija la orientación del hiperplano, mientras que el sesgo fija el desplazamiento del hiperplano en el espacio *d*-dimensional.

Hiperplano de separación

Un hiperplano divide el espacio original d-dimensional en dos semiespacios. Un grupo de datos es linealmente separable si cada semiespacio contiene puntos de una sola clase. Si el grupo de datos de entrada es linealmente separable, podemos encontrar un hiperplano de separación $h(\mathbf{x}) = 0$, tal que para todos los puntos etiquetados $y_i = -1$, tenemos $h(\mathbf{x}_i) < 0$, y para todos los puntos etiquetados $y_i = +1$, tenemos $h(\mathbf{x}_i) > 0$. De hecho, la función de hiperplano $h(\mathbf{x})$ sirve como clasificador lineal o un discriminante lineal, que predice la clase "y" para cualquier punto dado **x**, según la regla de decisión:

$$y = \begin{cases} +1 & si\ h(x) > 0 \\ -1 & si\ h(x) < 0 \end{cases}$$

Obsérvese que tanto **w** como -**w** son normales al hiperplano, eliminamos esta ambigüedad exigiendo que $h(\mathbf{x}_i) > 0$ cuando $y_i = 1$, y $h(\mathbf{x}_i) < 0$ cuando $y_i = -1$.

Distancia de un punto al hiperplano

Consideremos un punto $\mathbf{x} \in \mathbb{R}^d$ tal que **x** no se encuentra en el hiperplano. Sea $\mathbf{x}_p$ la proyección ortogonal de **x** sobre el hiperplano, y sea $\mathbf{r} = \mathbf{x} - \mathbf{x}_p$, entonces como se muestra en la función hiperplano podemos escribir **x** como

$$x = x_p + r$$

$$x = x_p + r\frac{w}{\|w\|}$$

donde *r* es la *distancia dirigida* del punto **x** desde $\mathbf{x}_p$, es decir, *r* da el desplazamiento de **x** desde $\mathbf{x}_p$ en términos del vector peso unitario $\frac{w}{\|w\|}$. El desplazamiento *r* es positivo si **r** está en la misma dirección que **w**, y *r* es negativo si **r** está en una dirección opuesta a **w**.

Introduciendo la ecuación anterior en la función hiperplano, obtenemos:

$$h(x) = h\left(x_p + r\frac{w}{\|w\|}\right)$$

$$= w^T\left(x_p + r\frac{w}{\|w\|}\right) + b$$

$$= w^T x_p + b + r\frac{w^T w}{\|w\|}$$

$$= h(x_p) + r\|w\|$$

$$= r\|w\|$$

El último paso se deduce del hecho de que $h(\mathbf{x}_p) = 0$ porque $\mathbf{x}_p$ se encuentra en el hiperplano. Utilizando el resultado anterior, obtenemos una expresión para la distancia dirigida de un punto al hiperplano:

$$r = \frac{h(x)}{\|w\|}$$

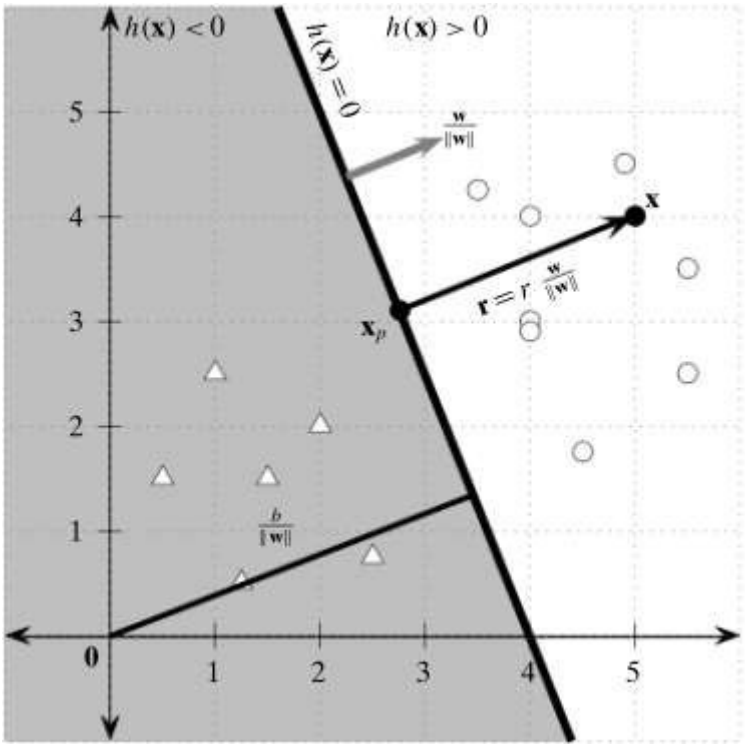


Figura 4. Geometría de un hiperplano de separación en 2D. Los puntos marcados +1 se muestran como círculos, y los marcados -1 se muestran como triángulos. El hiperplano $h(\mathbf{x})$ =0 divide el espacio en dos semi-espacios. La región sombreada comprende todos los puntos **x** que satisfacen $h(\mathbf{x}) < 0$, mientras que el área no sombreada está formada por los puntos que satisfacen $h(\mathbf{x}) > 0$. El vector peso unitario $\frac{w}{\|w\|}$ (en gris) es ortogonal al hiperplano. La distancia dirigida del origen al hiperplano es $\frac{b}{\|w\|}$.

Para obtener la distancia, que debe ser no negativa, podemos multiplicar convenientemente *r* por la etiqueta de clase *y* del punto, ya que cuando $h(\mathbf{x}) < 0$, la clase es -1, y cuando $h(\mathbf{x}) > 0$ la clase es +1. Así pues, la distancia de un punto **x** al hiperplano $h(\mathbf{x}) = 0$ es la siguiente:

$$\delta = yr = \frac{y\, h(x)}{\|w\|}$$

En particular, para el origen **x** = 0, la distancia dirigida es:

$$r = \frac{h(0)}{\|w\|} = \frac{w^T 0 + b}{\|w\|} = \frac{b}{\|w\|}$$

tal como se ilustra en la figura 4 (Mohammed & Wagner, 2020).

2.2.4. Regresión Logística Binaria

La regresión logística binaria presenta como característica más destacada el que la variable dependiente “Y” puede tomar solamente dos valores, éxito y fracaso (tipo binaria).

La regresión logística es un modelo estadístico empleado para estimar la probabilidad de que se produzca un suceso. El valor esperado de "y" se define como el intervalo entre 0 y 1, dadas las variables regresoras. Existe una relación no lineal entre las variables regresoras y la variable dependiente. Dado que las probabilidades estimadas estarán siempre comprendidas entre 0 y 1, la variable dependiente puede definirse como la probabilidad de que se produzca o no un acontecimiento controlable.

El proceso de regresión logística implica la selección de los coeficientes del modelo que producen los resultados más "probables". Para estimar los parámetros de la regresión logística es necesario un algoritmo iterativo debido a su naturaleza no lineal.

Métodos de estimación

El modelo a encontrar haciendo uso de la regresión logística puede ser estimado utilizando la metodología de máxima verosimilitud (MV) o la metodología modelo lineales generalizados (GLM). La estimación de máxima verosimilitud usa típicamente formas modificadas de ecuaciones de estimación de Newton-Raphson; GLM utiliza un algoritmo de mínimos cuadrados reponderados iterativamente (IRLS) que es una simplificación de la estimación de máxima verosimilitud, pero que se limita a distribuciones que pertenecen a la familia exponencial de distribuciones. En el caso de máxima verosimilitud, una ecuación de estimación se define como establecer en 0 la derivada de la función logarítmica de verosimilitud de la distribución de respuesta con respecto a uno de los parámetros de interés, donde hay una única ecuación de estimación para cada parámetro desconocido. El término ecuación de estimación se conoce como tal porque su solución da como resultado estimaciones puntuales de sus diversos parámetros. A continuación, discutimos la derivación de las ecuaciones de estimación de máxima verosimilitud en las que se basan los modelos de regresión logística junto con los errores estándar de las estimaciones y variedades de estimación de MV.

Los errores estándar (EE) se basan en la matriz hessiana negativa inversa. Tradicionalmente, el método utilizado para modelar GLM emplea el hessiano como

matriz de información esperada (EIM). En general, esto es mucho más simple que usar la matriz de información completa, también llamada matriz de información observada (OIM). La única diferencia entre los dos se relaciona con los errores estándar en modelos con pocas observaciones. OIM es preferible y se utiliza en la estimación de MV regular. El algoritmo GLM se puede modificar para utilizar OIM para todos los enlaces, canónicos y no canónicos por igual. Sin embargo, debe tenerse en cuenta que para un modelo vinculado canónico, el OIM se reduce al EIM.

Derivación del Algoritmo IRLS

La estimación se basa en una modificación de una expansión de Taylor de dos términos de la función logarítmica de verosimilitud. En su forma original, la expansión de Taylor aparece como:

$$0 = f(X_0) + (X_1 - X_0) + f'(X_0) + \frac{(X_1-X_0)^2}{2!} f''(X_0) + \frac{(X_1-X_0)^3}{3!}\ f'''(X_0) + \ldots$$

Los dos primeros términos se reducen a:

$$0 = f(X_0) + (X_1 - X_0) + f'(X_0)$$

que se puede refundir para:

$$X_1 = X_0 - \frac{f(X_0)}{f'(X_0)}$$

El método de estimación de Newton-Raphson adopta lo anterior utilizando la puntuación o el gradiente de la función logarítmica de verosimilitud como base de la estimación de parámetros. La forma es:

$$\beta_r = \beta_{r-1} - \frac{\partial L\beta_{r-1}}{\partial^2 L\beta_{r-1}}$$

Dónde:

$$\partial L = \frac{\partial L}{\partial \beta}$$

y

$$\partial^2 L = \frac{\partial^2 L}{\partial \beta \partial \beta'}$$

En una forma más tradicional, sea:

$$\mu = \partial L \;\; y \;\; H = \partial^2 L$$

entonces,

$$\beta_r = \beta_{r-1} - H^{-1}\mu$$

donde,

$$H = H_{r-1} \;\; y \;\; \mu = \mu_{r-1}$$

Resolver para ∂L o U: el gradiente

En forma de familia exponencial, la función logarítmica de verosimilitud se expresa como:

$$\frac{\partial L}{\partial \beta} = \sum \frac{y\theta - b(\theta)}{\propto (\emptyset)} + \sum C(y, \emptyset)$$

Por la regla de la cadena:

$$\frac{\partial L}{\partial \beta} = \frac{\partial L}{\partial \theta} X \frac{\partial \theta}{\partial \mu} X \frac{\partial \mu}{\partial \eta} X \frac{\partial \eta}{\partial \beta}$$

Resolviendo para cada término:

$$\frac{\partial L}{\partial \beta} = \sum \frac{y\theta - b'(\theta)}{\propto (\emptyset)} + \sum \frac{y - \mu}{\alpha(\emptyset)}$$

ya que:

$$b'(\theta) = \mu$$

$$\frac{\partial \mu}{\partial \theta} = \frac{\partial b'(\theta)}{\partial \theta} = b''(\theta) = V(\mu)$$

$$\frac{\partial \theta}{\partial \mu} = \frac{1}{V(\mu)}$$

$$\frac{\partial \eta}{\partial \beta} = \frac{\partial x\beta}{\partial \beta}, ya\ que\ \eta = x\beta$$

$$\frac{\partial \eta}{\partial \beta} = [g^{-1}(\eta)]' = \frac{1}{\partial \eta / \partial \mu} = \frac{1}{g'(\mu)}$$

que es la derivada de la función de enlace wrt μ.

La sustitución de expresiones da a los estimadores de máxima verosimilitud de b la solución de la ecuación de estimación basada en vectores:

$$\sum \frac{(y-\mu)x}{\alpha(\emptyset)V(\mu)g'(\mu)} = \frac{(y-\mu)x}{\alpha(\emptyset)V(\mu)}\frac{\partial\mu}{\partial\eta} = 0$$

donde "y" y "μ" son escalares, x es un vector de fila de 1 x p, y la suma resultante es un vector de columna de p x 1 de 0s.

Resolviendo para d^2L - Puntaje de Fisher

El algoritmo GLM tradicional sustituye a I por H, la matriz hessiana de segundas derivadas observadas. I es la segunda de dos formas equivalentes de información de Fisher dada por:

$$I = -E\left[\frac{\partial^2 L}{\partial\beta\partial\beta'}\right] = E\left[\frac{\partial L}{\partial\beta}\frac{\partial L}{\partial\beta'}\right]$$

Resolviendo los rendimientos anteriores:

$$I = \frac{\partial}{\partial\beta_j}\left[\frac{(y-\mu)x_j}{\alpha(\emptyset)V(\mu)}\frac{\partial\mu}{\partial\eta}\right] * \frac{\partial L}{\partial\beta_k}\left[\frac{(y-\mu)x_k}{\alpha(\emptyset)V(\mu)}\frac{\partial\mu}{\partial\eta}\right]$$

$$I = \frac{(y-\mu)^2 x_j x_k}{\alpha(\emptyset)V(\mu)^2}\left[\frac{\partial\mu}{\partial\eta}\right]^2$$

ya que:

$$(y-\mu)^2 = \alpha(\emptyset)V(\mu)$$

y dejando:

$$V(y) = \alpha(\emptyset)V(\mu) = (y-\mu)^2$$

Volviendo a formular como:

$$I = \frac{x_j x_k}{V(y)}\left[\frac{\partial\mu}{\partial\eta}\right]^2 = \frac{x_j x_k}{V(y)g'^2}$$

Juntando las diversas ecuaciones:

$$\beta_r = \beta_{r-1} = \left[\frac{x_j x_k}{V(y)}\left[\frac{\partial\mu}{\partial\eta}\right]^2\right]^{-1}\left[\frac{x_k(y-\mu)}{V(y)}\left[\frac{\partial\mu}{\partial\eta}\right]\right]$$

Multiplica ambos lados por I:

$$\left[\frac{x_j x_k}{V(y)}\left[\frac{\partial\mu}{\partial\eta}\right]^2\right]\beta_r = \left[\frac{x_j x_k}{V(y)}\left[\frac{\partial\mu}{\partial\eta}\right]^2\right]\beta_{r-1} + \left[\frac{x_k(y-\mu)}{V(y)}\left[\frac{\partial\mu}{\partial\eta}\right]\right]$$

Definiendo W:

$$W(weight) = \frac{1}{V(y)}\left[\frac{\partial\mu}{\partial\eta}\right]^2$$

y

$$\eta = x_k\beta_{r-1}$$

Convierta la representación algebraica anterior a forma matricial. Esto se puede hacer por partes. Primero, dada la definición de W anterior, se puede hacer la siguiente sustitución:

$$\left[\frac{x_j x_k}{V(y)}\left[\frac{\partial\mu}{\partial\eta}\right]^2\right]\beta_r = [X'WX]\beta_r$$

En segundo lugar, recordando la definición de V(y) y W:

$$\frac{x_k(y-\mu)}{V(y)}\left[\frac{\partial\mu}{\partial\eta}\right] = \frac{x_k(y-\mu)}{\frac{1}{w}\left[\frac{\partial\mu}{\partial\eta}\right]^2}\left[\frac{\partial\mu}{\partial\eta}\right]$$

Tercero, ya que:

$$\eta = x_k\beta_{r-1}$$

$$\left[\frac{x_j x_k}{V(y)}\left[\frac{\partial\mu}{\partial\eta}\right]^2\right]\beta_{r-1} = X'W\eta$$

Combinando términos, tenemos:

$$[X'WX]\beta_r = X'W\eta + \left[\frac{x_k(y-\mu)}{\frac{1}{w}\left[\frac{\partial\mu}{\partial\eta}\right]^2}\left[\frac{\partial\mu}{\partial\eta}\right]\right]$$

$$[X'WX]\beta_r = X'W\eta + \left[X_k W(y-\mu)\left(\frac{\partial\eta}{\partial\mu}\right)\right]$$

Dejamos:

$$z = \eta + (y - \mu)\left(\frac{\partial \eta}{\partial \mu}\right)$$

$$[X'WX]\beta_r = X'Wz$$

así que:

$$\beta_r = [X'WX]^{-1}X'Wz$$

que es una matriz de regresión ponderada que se utiliza para actualizar iterativamente las estimaciones del vector de parámetros β_r, así como los valores de μ y η, y la función de desviación, que se define mediante las ecuaciones:

$$D = 2\sum\{LL(y; y) - LL(\mu; y)\}$$

que se calcula en:

$$D = 2\sum\left\{yln\left(\frac{y}{\mu}\right) + (1 - y)\ln(1 - y)\,/(1 - \mu))\right\}$$

El umbral de iteración para software comercial generalmente se establece en 10^{-6}.

El umbral de convergencia se basa en la minimización de las diferencias entre las estadísticas antiguas y nuevas de desviación o log-verosimilitud, o en los valores respectivos dentro del vector de estimaciones de parámetros. Tradicionalmente, el software GLM repetía la estadística de desviación y luego calculaba la probabilidad logarítmica para su uso en varias estadísticas de bondad de ajuste (Hilbe, 2009).

Estimación máxima verosimilitud

Los procedimientos de estimación de máxima verosimilitud (MLE, por sus siglas en inglés), o ML, suelen utilizar una variedad de algoritmos de Newton-Raphson modificados. La metodología de máxima verosimilitud completa utiliza la matriz de información observada en la que se basan los errores estándar del modelo. Vimos como Fisher Scoring permite una simplificación de esta matriz en el caso de distribuciones pertenecientes a la familia exponencial. En aras de la exhaustividad, describimos las fórmulas utilizadas para el cálculo de la matriz de información observada y mostramos cómo se puede reducir cuando se aplica a modelos de familias exponenciales como GLM.

La derivación de términos para el algoritmo de estimación comienza como una linealización de Taylor y continúa hasta el cálculo del gradiente, o primera derivada de la

función de verosimilitud. Fisher Scoring, utilizado como base del algoritmo de estimación GLM, calcula la matriz de segundas derivadas con base en la matriz de información esperada (EIM), o I, donde:

$$I = -E\left[\frac{\partial^2 L}{\partial\beta\partial\beta'}\right] = E\left[\frac{\partial L}{\partial\beta}\frac{\partial L}{\partial\beta'}\right]$$

La metodología de Newton-Raphson, o MLE completa, calcula las segundas derivadas de la probabilidad sobre la base de la matriz de información observada (OIM), que permite la estimación de algoritmos basados en la probabilidad distintos de la forma de familia exponencial más limitada. Nos referimos a esta matriz como H, que significa Hessian.

$$H = \left[\frac{\partial^2 L}{\partial\beta_j\partial\beta_k}\right] = \sum_{i=1}^{n}\frac{1}{\propto(\theta)}\left[\frac{\partial}{\partial\beta_k}\right]\left\{\frac{y_i-\mu_i}{V(\mu_i)}\left(\frac{\partial\mu}{\partial\eta}\right)_i x_j x_i\right\}$$

Resuelto, lo anterior se convierte en:

$$-\sum_{i=1}^{n}\frac{1}{\alpha(\phi)}\left[\frac{1}{V(\mu_i)}\left(\frac{\partial\mu}{\partial\eta}\right)_i^2-(\mu_i-y_i)\left\{\frac{1}{V(\mu_i)^2}\left(\frac{\partial\mu}{\partial\eta}\right)_i^2\frac{\partial V(\mu_i)}{\partial\mu}-\frac{1}{V(\mu_i)}\left(\frac{\partial^2\mu}{\partial\eta^2}\right)_i\right\}\right]x_{ji}x_{ki}$$

Tenga en cuenta que en el caso del enlace canónico, los términos de −(μ − y) anteriores se cancelan, reduciéndose al valor del EIM. Esto sucede con el algoritmo logístico, así como con las formas canónicas de todos los modelos GLM. Esquematizando lo anterior a un algoritmo genérico de tipo Newton-Raphson nos da el algoritmo MLE.

Las estimaciones de los parámetros, los errores estándar, etc. del modelo de regresión logística, al ser la forma canónica de la familia binomial, son idénticos independientemente del método de estimación, es decir, máxima verosimilitud total (ML) o mínimos cuadrados reponderados iterativos (IRLS). Debe tenerse en cuenta que la estimación de IRLS también se conoce como puntuación de Fisher.

En resumen, dado que el modelo logístico es la forma canónica de la familia binomial general de modelos, la matriz de información observada más complicada utilizada para calcular los errores estándar se reduce a la matriz de información esperada. La matriz de información esperada se usa solo en modelos lineales generalizados (GLM). Por lo tanto, se deduce que el modelo logístico se puede estimar utilizando la metodología GLM o MLE. Ambos métodos producen resultados idénticos. Muchos estadísticos

prefieren usar el enfoque GLM, sin embargo, debido a la gran cantidad de estadísticas de ajuste que normalmente se asocian con el procedimiento tal como se implementa en el software comercial.

Derivación del algoritmo logístico binario

El modelo logístico se deriva de la función de distribución de probabilidad (PDF) de Bernoulli, que es un subconjunto de la binomial general.

p ó π, son términos estándar que se utilizan para indicar la probabilidad de éxito, que van desde cerca de 0 hasta cerca de 1. Preferiremos p en lugar de π. La variable *y* generalmente representa la respuesta real del modelo, con un valor de 1 ó 0. Un valor de y = 1 indica un éxito o satisfacción de un criterio dado; y = 0 representa una falta de éxito, lo que puede considerarse como una falla en satisfacer el criterio definido. Dados estos términos, la PDF de Bernoulli se puede expresar, para una sola observación, como:

PDF Bernoulli

$$f(y_i; p_i) = \prod_{i=1}^{n} {p_i}^{y_i} \, (1 - p_i)^{1-y_i}$$

expresado en términos de la forma de familia exponencial, sin subíndices:

$$f(y; \theta; \phi) = \prod \exp \{\frac{y\theta - b(\theta)}{\alpha(\phi)} + c(y; \phi)\}$$

tenemos:

$$f(y; p) = \prod \exp \{yln(\frac{p}{1-p}) + ln(1-p)\}$$

La ecuación anterior se puede volver a parametrizar para estimar *p* dado *y*, en lugar de lo contrario, como es el caso de las funciones de probabilidad. La función resultante se llama función de verosimilitud y se presenta como:

$$L(p; y) = \prod \exp \{yln(\frac{p}{1-p}) + ln(1-p)\}$$

La familia de distribuciones exponenciales de un solo parámetro tiene una característica única y extremadamente útil. Si se parametriza una distribución para reflejar la estructura anterior, el enlace, la varianza y la media de la distribución se pueden calcular en función de las siguientes relaciones.

Enlace

$$\theta = ln\left(\frac{p}{1-p}\right) = \eta = \ln\left(\frac{\mu}{1-\mu}\right)$$

Media

La media se calcula como la primera derivada del cumulante, b(θ), wrt θ:

$$b(\theta) = -\ln(1-p)$$

$$b'(\theta) = p = \mu$$

donde b′(θ) es la derivada de b(θ) con respecto a θ. Observe que ahora podemos definir μ como idéntico a p. μ es el término tradicional utilizado en modelos lineales generalizados para la media. En el algoritmo GLM real, μ, el valor ajustado, se calcula a partir del predictor lineal como el enlace inverso.

Inversa del enlace

$$\mu = g^{-1}(\eta)$$

$$= 1/(1+e^{-\eta})$$

ó

$$= e^{\eta}/(1+e^{\eta})$$

El enlace, definido de esta manera, generalmente se representa como g(μ). Por lo tanto, tenemos:

$$g(\mu) = \ln\left(\frac{\mu}{1-\mu}\right)$$

y la derivada de $g(\mu)$ como

$$g'(\mu) = 1/\left(\frac{\mu}{1-\mu}\right)$$

Varianza

La varianza es calculada como la segunda derivada de b(θ), wrt θ:

$$b''(\theta) = p(1-p) = \mu(1-\mu)$$

que es simplemente la segunda derivada del cumulante.

Log-Verosimilitud

El log-verosimilitud se determina tomando el logaritmo natural de la función de verosimilitud para que aparezca en forma de familia exponencial como:

$$L(\mu; y) = \sum \left\{ y ln \left(\frac{\mu}{1-\mu} \right) + \ln(1-\mu) \right\}$$

Tenga en cuenta la suma de las observaciones en todo el modelo en lugar del producto empleado. El registro de la función permite la suma, que es mucho más fácil de usar en los cálculos que la multiplicación. Por esta razón, el log-verosimilitud siempre se usa con preferencia a la verosimilitud para fines de maximización.

Muchas formulaciones de la función log-verosimilitud, cuando se utilizan como medida de convergencia en una regresión logística binaria, se dividen en dos funciones: una cuando y = 1 y la otra cuando y = 0. Esta representación dual de la log-verosimilitud aparece como:

$$L(y = 1) = \sum \left\{ ln \left(\frac{\mu}{1-\mu} \right) + \ln\,(1-\mu) \right\}$$

Y

$$L(y = 0) = \sum \ln(1-\mu)$$

Dado que el predictor lineal logístico, *xb*, es η, podemos parametrizar la log-verosimilitud en términos de *xb* en lugar de μ. De hecho, los algoritmos de máxima verosimilitud requieren una estimación en términos de *xb*. Hemos visto anteriormente que la relación de *xb* con μ es

$$\mu = \exp(xb)\,/(1 + \exp(xb))$$

Como consecuencia, la conversión algebraica da como resultado que la función log-verosimilitud logística binaria se exprese como:

$$L(xb; y) = \sum \{ y(xb) - ln(1 + \exp(xb)) \}$$

Lo anterior también se observa de la siguiente forma, utilizando β en lugar de b, como:

$$L(x\beta; y) = \sum \{ y(x\beta) - ln(1 + \exp(x\beta)) \}$$

Definición

Para los problemas de clasificación necesitamos algo mejor que la función lineal. Tendemos a utilizar una función sigmoidal, como la función logística,

$$h_\beta(x) = \frac{1}{1 + e^{-\beta_0 - \beta_1 x}}$$

o, ya que vamos a pasar inmediatamente a la regresión múltiple,

$$h_\beta(x) = \frac{1}{1 + e^{-\beta^T x}}$$

donde los β_s son los parámetros que queremos encontrar para que nos den el mejor ajuste a los datos.

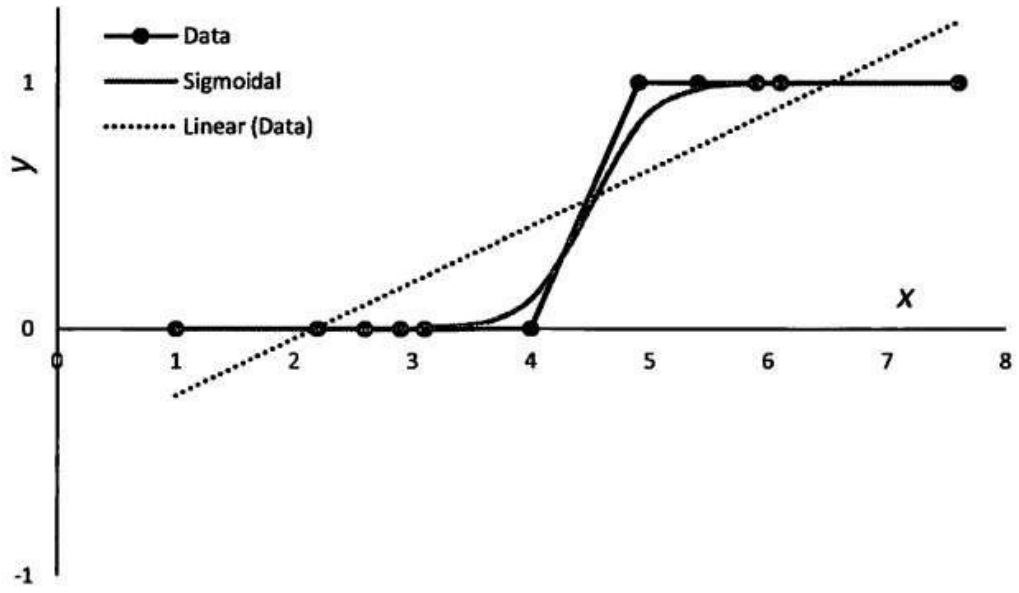


Figura 5. Regresión para un problema de clasificación.

Ajustaríamos esta función a los datos y, a continuación, dado un nuevo punto de datos, determinaríamos si presenta la característica o no, según un umbral para h_β. Podríamos tener un umbral de 0,5, de modo que cualquier cosa por encima de ese nivel iría a nuestra bandeja que presenta la característica. O podríamos utilizar un umbral diferente (Wilmott, 2019).

La función de coste

Una diferencia importante cuando queremos hacer regresión logística en lugar de regresión lineal es la elección de la función de coste.

Existen algunas propiedades obvias para una función de coste: debe ser positiva excepto cuando tenemos un ajuste perfecto, en cuyo caso debe ser cero; debe tener un único mínimo. Nuestra función de costes cuadrática, cumple la primera de estas propiedades, pero cuando h_β es la función logística no tiene por qué tener un único mínimo.

Sin embargo, la siguiente función de coste sí tiene estas propiedades.

$$J(\beta) = -\frac{1}{N}\sum_{n=1}^{N}\left(y^{(n)}\ln\left(h_\beta(x^{(n)})\right) + (1-y^{(n)})\ln\left(1-h_\beta(x^{(n)})\right)\right)$$

¿Por qué funciona esta función de costes? Recordemos que "y" sólo puede tomar valores 0 ó 1. Cuando y = 1 la función $J(\beta)$ decrece suave y monótonamente hasta un valor de cero cuando h_β también es uno. Y cuando y = 0, la función de coste también es cero cuando h_β = 0. Pero quizás lo más importante es que esta función de coste tiene una buena interpretación en términos de estimación de máxima verosimilitud para problemas de clasificación.

Ya no tenemos una solución analítica para el mínimo de la función de coste, así que tenemos que resolver numéricamente. Pero lo más bonito es que con la nueva definición de h_β y la nueva función de coste seguimos teniendo

$$\frac{\partial J}{\partial \beta} = \frac{1}{N}\sum_{n=1}^{N} x^{(n)}\left(h_\beta\left(x^{(n)}\right) - y^{(n)}\right)$$

Esto significa que nuestro algoritmo de descenso gradiente permanece, sorprendentemente, sin cambios (Wilmott, 2019).

2.2.5. Árbol de decisión

El árbol está dibujado al revés, con la raíz (la primera pregunta) en la parte superior. Cada pregunta es una condición o un nodo interno que divide características o atributos. De cada nodo saldrán ramas o aristas, que representan las posibles respuestas. Cuando se llega al final del camino a través del árbol de modo que no hay más preguntas / decisiones entonces usted está en una hoja. También existen los conceptos obvios de ramas y nodos padre e hijo (Wilmott, 2019). Véase la Figura 6.

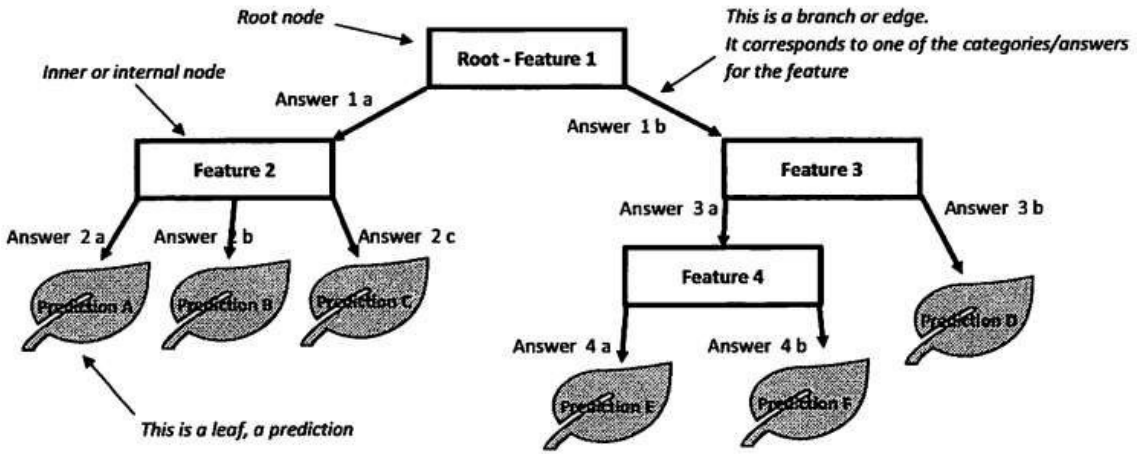


Figura 6. Esquema de las características de un árbol de decisión.

Se puede utilizar árboles de decisión para clasificar datos, como si una seta es comestible o no, teniendo en cuenta varias características físicas. Estas características pueden tener categorías binarias, como con o sin branquias, o múltiples categorías.

Construir o hacer crecer un árbol consiste en elegir el orden en que se examinan las características de los datos y cuáles son las condiciones. Es posible que haya que hacer alguna poda y habrá que saber cuándo parar.

El algoritmo de los árboles de decisión

Paso 1: Elige un atributo

Tome uno de los atributos y ramas (restantes) y divida los datos.

Paso 2: Calcular la entropía

Calcule la entropía de cada rama. A continuación, calcule la entropía media de ambas/todas las ramas para ese atributo.

Vuelva al paso 1 hasta haber examinado todos los atributos.

Paso 3: Establecer el atributo que minimiza la entropía

Elija como nodo el atributo que minimice la entropía (o, lo que es lo mismo, que maximice la ganancia de información en no divididos).

Desplácese hacia abajo/atraviese el árbol y vuelva al Paso 1.

2.2.6. AdaBoost

El boosting utiliza subconjuntos aleatorios de ejemplos de entrenamiento extraídos del grupo de datos de entrenamiento sin sustitución.

A diferencia del procedimiento de Boosting, AdaBoost utiliza el grupo de datos de entrenamiento completo para entrenar a los "aprendices débiles" (un ejemplo típico de un aprendiz débil es un árbol de decisión de un nivel), donde los ejemplos de entrenamiento se reponen en cada iteración para elaborar un clasificador fuerte que aprende de los errores de los aprendices débiles anteriores en el conjunto.

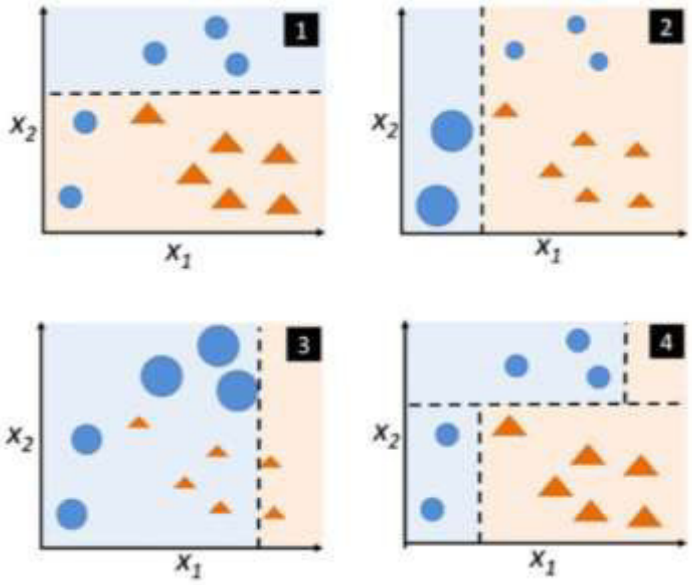


Figura 7. Concepto de AdaBoost para mejorar a los aprendices débiles.

Para recorrer la ilustración de AdaBoost paso a paso, con la figura 7, que representa un grupo de datos de entrenamiento para clasificación binaria en el que se asignan pesos iguales a todos los ejemplos de entrenamiento. Basándonos en este grupo de datos de entrenamiento, entrenamos una decisión de un nivel (mostrado como una línea discontinua) que intenta clasificar los ejemplos de las dos clases (triángulos y

círculos), así como posiblemente minimizar la función de pérdida (o la puntuación de impureza en el caso especial de los conjuntos de árboles de decisión).

Para la figura 7, asignamos un mayor peso a los dos ejemplos previamente mal clasificados (círculos). Además, reducimos el peso de los ejemplos correctamente clasificados. La siguiente decisión se centrará ahora en los ejemplos de entrenamiento que tienen los pesos más grandes, los ejemplos de entrenamiento que supuestamente son difíciles de clasificar.

El aprendiz débil mostrado en la figura 7 clasifica erróneamente tres ejemplos diferentes de la clase círculo, a los que se les asigna un peso mayor, como se muestra en la figura 3.

Suponiendo que nuestro conjunto AdaBoost solo consta de tres rondas de boosting, combinamos los tres aprendices débiles entrenamos en diferentes subconjuntos de entrenamientos reponderados mediante un voto mayoritario ponderado, como se observa en la figura 7.

Algoritmo utilizando un pseudocódigo

1. Se define el vector de pesos, w, con pesos uniformes, donde $\sum_i w_i$=1.
2. Para j en m rondas de refuerzo, se hace lo siguiente:
    a. Se entrena un aprendiz débil ponderado: $C_j = train(X, y, w)$.
    b. Se pronostican las etiquetas de clase: $\hat{y} = predict(C_j, X)$.
    c. Se calcula la tasa de error ponderada: $\varepsilon = w \cdot (\hat{y} \neq y)$.
    d. Se calcula el coeficiente: $\alpha_j = 0.5 \log \frac{1-\varepsilon}{\varepsilon}$.
    e. Se actualizan los pesos: $w := w \times \exp(-\alpha \times \hat{y} \times y)$.
    f. Se normalizan los pesos para que sumen 1: $w := w / \sum_i w_i$.
3. Se calcula el pronóstico final: $\hat{y} = \left(\sum_{j=1}^{m} \left(\propto_j \times predict(C_j, X)\right) > 0\right)$.

(Raschka et al., 2023).

2.2.7. Gradient Boosting

El Gradient Boosting es muy similar al AdaBoost. AdaBoost entrena árboles de decisión de un solo nivel basados en errores del anterior árbol de decisión de un solo nivel. En concreto, los errores se utilizan para calcular los pesos de las muestras en cada ronda, así como para calcular el peso del clasificador para el árbol de decisión de un nivel

cuando se combinan los niveles individuales en un conjunto. Dejamos de entrenar cuando se logra el número máximo de iteraciones (árboles de decisión de un nivel). Al igual que AdaBoost, el Gradient Boosting ajusta los árboles de decisión de forma iterativa utilizando los errores de pronóstico. Sin embargo, los árboles de Gradient Boosting suelen ser más profundos que los árboles de decisión y suelen tener una profundidad máxima de 3 a 6 (o un número máximo de 8 a 64 nodos hoja). Además, a diferencia de AdaBoost, el Gradient Boosting no utiliza los errores de pronóstico para asignar pesos a las muestras; se utilizan directamente para formar la variable objetivo para ajustar el siguiente árbol. Además, en lugar de tener un término de ponderación individual para cada árbol, como el AdaBoost, el Gradient Boosting utiliza una tasa de aprendizaje global que es la misma para cada árbol (Raschka et al., 2023).

Algoritmo de Gradient Boosting para la clasificación binaria

Utilizaremos la función de pérdida logística. Para un solo ejemplo de entrenamiento, podemos especificar la pérdida logística como sigue:

$$L_i = -y \log p_i + (1 - y) \log(1 - p)$$

Log(odds):

$$\hat{y} = \log(odds) = log\left(\frac{p}{1-p}\right)$$

Se utiliza log(odds) para reescribir la función logística como sigue:

$$L_i = \log\left(1 + e^{\hat{y}_i}\right) - y\,\hat{y}_i$$

Ahora, se define la derivada parcial de la función de pérdida con respecto a estos log(odds), $\hat{y}$. La derivada de esta función de pérdida con respecto a log(odds) es:

$$\frac{\partial L_i}{\partial \hat{y}_i} = \frac{e^{\hat{y}_i}}{1 + e^{\hat{y}_i}} - y_i = p_i - y_i$$

Pasos generales de Gradient Boosting

1. Crear un nodo raíz que minimice la pérdida logística. Resulta que la pérdida se minimiza si el nodo raíz devuelve el log(odds), $\hat{y}$.
2. Para cada árbol m=1, …, M, donde M es un número de árboles totales especificado por el usuario, llevar a cabo los siguientes cálculos descritos en los pasos del 2a al 2d:
   a. Convertir log(odds) en una probabilidad utilizando la conocida función logística

que se utiliza en la regresión logística:

$$p = \frac{1}{1 + e^{-\hat{y}}}$$

A continuación, calcular el pseudoresiduo, que es la derivada parcial negativa de la pérdida con respecto a log(odds), que resulta ser la diferencia entre la etiqueta de clase y la probabilidad pronosticada:

$$-\frac{\partial L_i}{\partial \hat{y}_i} = y_i - p_i$$

b. Adaptar un nuevo árbol a los pseudoresiduos.

c. Para cada nodo R_{jm}, calcular el valor γ_{jm} que minimice la función de pérdida logística. Esto incluye una fase de síntesis para tratar los nodos hoja que contienen múltiples ejemplos de entrenamiento:

$$\gamma_{jm} = \underset{\gamma}{\arg min} \sum_{x_i \in R_{jm}} L(y_i, F_{m-1}(x_i) + \gamma)$$

$$= \log\left(1 + e^{\hat{y}_i + \gamma}\right) - y_i(\hat{y}_i + \gamma)$$

Omitiendo los detalles matemáticos intermedios, se obtiene lo siguiente:

$$\gamma_{jm} = \frac{\sum_i y_i - p_i}{\sum_i p_i(i - p_i)}$$

Hay que tener en cuenta que la suma aquí es solo sobre los ejemplos en el nodo correspondiente al nodo hoja R_{jm} y no el conjunto de entrenamiento completo.

d. Actualizar el modelo añadiendo el valor gamma del paso 2c con la tasa de aprendizaje η:

$$F_m(x) = F_{m-1}(x) + \eta\gamma_m$$

(Raschka et al., 2023).

2.2.8. XGBoost

XGBoost propuso varios trucos y aproximaciones que aceleran sustancialmente el proceso de entrenamiento. De ahí el nombre XGBoost, que significa eXtreme Gradient Boosting (refuerzo del gradiente extremo).

XGBoost es una librería distribuida de gradient boosting que ha sido optimizada para su **portabilidad, flexibilidad y eficiencia**. El sistema ejecuta algoritmos de machine learning utilizando la arquitectura Gradient Boosting. XGBoost ofrece boosting de árbol

paralelo (GBDT, GBM), que resuelve de forma eficiente y precisa numerosos retos de la ciencia de datos (https://xgboost.readthedocs.io/en/stable/).

2.2.9. Bagging

El método de agregación Bootstrap o "bagging" es un método en el que se seleccionan muestras con reemplazo, de un árbol completo o de un conjunto de datos completo y luego se aplica la técnica de machine learning que se esté utilizando para obtener un pronóstico, y luego se repite con otra muestra aleatoria. Se hace esto muchas veces y la previsión final en un problema de clasificación es el voto mayoritario (Wilmott, 2019).

El concepto de bagging se resume en la figura 8 (Raschka et al., 2023).

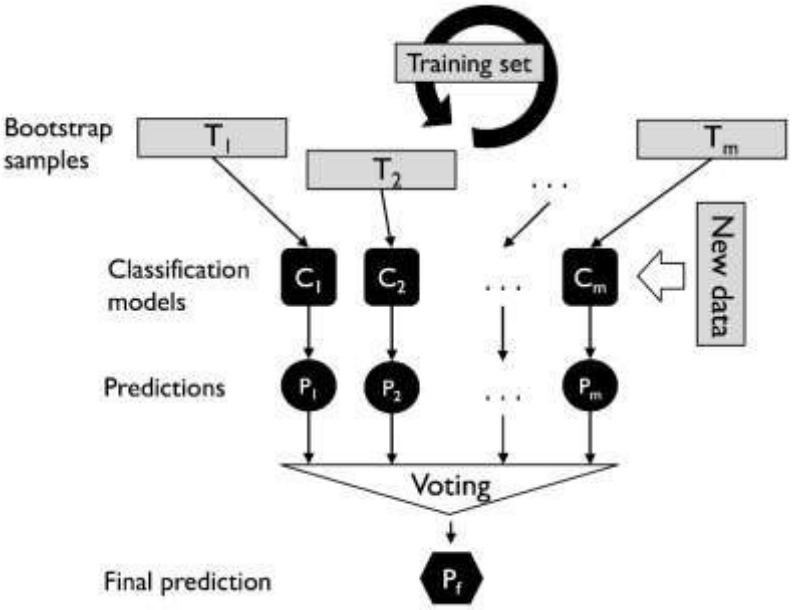


Figura 8. Concepto de bagging.

2.2.10. Random Forest

Random forest puede considerarse un grupo de arboles de decisión. La idea que hay detrás del random forest es promediar múltiples arboles de decisión (profundos) que individualmente reportan una alta variabilidad para construir un modelo que tenga un mayor rendimiento de generalización y sea menos susceptible al sobreajuste. El algoritmo del random forest puede resumirse en cuatro sencillos pasos:

1. Extraemos una muestra bootstrap aleatoria de tamaño *n* (elegimos aleatoriamente *n* ejemplos del conjunto de datos de entrenamiento con sustitución).

2. Hacemos crecer un árbol de decisión a partir de la muestra bootstrap. En cada nodo:
   a. Seleccionamos aleatoriamente *d* características sin sustitución.
   b. Dividimos el nodo utilizando la característica que proporciona la mejor división según la función objetivo; por ejemplo, maximizando la ganancia de información.
3. Repetimos los pasos de 1000 a 2000 veces.
4. Agregamos el pronóstico de cada árbol para asignar la etiqueta de clase por mayoría de votos.

Es importante destacar una pequeña modificación en la fase 2 del entrenamiento del árbol de decisión individual: en lugar de evaluar cada característica para determinar la división óptima en cada nodo, sólo tenemos en cuenta un subconjunto aleatorio.

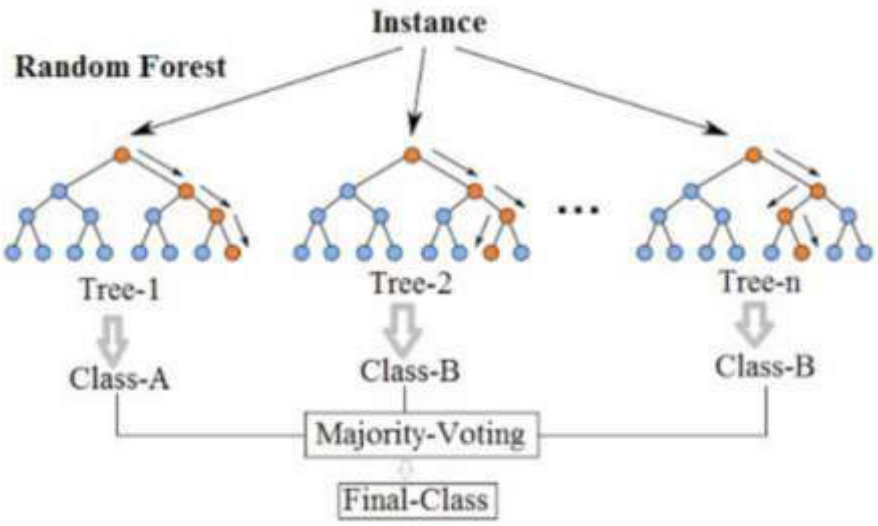


Figura 9. Concepto de Random Forest.

Aunque Random Forest no ofrecen el mismo nivel de interpretabilidad que los árboles de decisión, una gran ventaja del Random Forest es que no tenemos que preocuparnos tanto por elegir buenos valores de los hiperparámetros. Normalmente no necesitamos podar el Random Forest, ya que el modelo de conjunto es bastante robusto al ruido que se produce al promediar los pronósticos entre los árboles de decisión individuales. El único parámetro que debemos tener en cuenta en la práctica es el número de árboles, *k*, (paso 3) que elegimos para el Random forest. Normalmente, cuanto mayor sea el número de árboles, mejor será el rendimiento del clasificador de Random Forest, a costa de un mayor coste computacional (Raschka et al., 2023).

2.2.11. ExtraTrees

Al igual que Random Forest, ExtraTrees es un enfoque de machine learning por conjuntos que entrena numerosos árboles de decisión y agrega los resultados del grupo de árboles de decisión para generar una predicción. Sin embargo, hay pocas diferencias entre Extra Trees y Random Forest.

Random Forest utiliza el embolsado para seleccionar diferentes variaciones de los datos de entrenamiento para garantizar que los árboles de decisión sean lo suficientemente diferentes. Sin embargo, Extra Trees utiliza toda la base de datos para entrenar árboles de decisión. Por lo tanto, para garantizar unas disimilitudes adecuadas entre los árboles de decisión individuales, se hace una selección aleatoria de los valores que se utilizarán para dividir una entidad y generar nodos secundarios.

El uso de todo el conjunto de datos (que es la configuración predeterminada y se puede cambiar) permite a ExtraTrees reducir el sesgo del modelo. Sin embargo, la aleatorización del valor de la característica en la que se va a dividir aumenta el sesgo y la varianza.

ExtraTrees Classifier es otra de las técnicas de machine learning basada en árboles de decisión que utiliza la aleatorización para reducir la varianza y el costo computacional (en comparación con Random Forest) (Thankachan, 2022).

2.2.12 Evaluación de modelos de clasificación

La matriz de confusión determina el rendimiento de una técnica de aprendizaje, es una matriz cuadrada que muestra los recuentos de los pronósticos verdaderos positivos (True Positive, TP), verdaderos negativos (True Negative, TN), falsos positivos (False Positive, FP) y falsos negativos (False Negative, FN) de un clasificador, como se observa en la figura 10.

		Clase pronosticada	
		P	N
Clase real	P	Verdaderos positivos (TP)	Falsos negativos (FN)
	N	Falsos positivos (FP)	Verdaderos negativos (TN)

Figura 10. Matriz de confusión.

Optimización de la precisión y la recuperación

Tanto la precisión (ACC) como el error de pronóstico (ERR), proporcionan información general sobre el número de ejemplos mal clasificados. La precisión se calcula como la suma de pronósticos correctos dividida por el total de pronósticos y el error puede entenderse como la suma de todos los pronósticos falsos dividido por el número de pronósticos totales, respectivamente:

$$ERR = \frac{FP + FN}{FP + FN + TP + TN}$$

La precisión del pronóstico (Accuracy) puede calcularse directamente a partir del error:

$$ACC = \frac{TP + TN}{FP + FN + TP + TN} = 1 - ERR$$

La tasa de falsos positivos (FPR) y la tasa de verdaderos positivos (TPR) son métricas de rendimiento especialmente útiles para los problemas de clases desequilibradas:

$$FPR = \frac{FP}{N} = \frac{FP}{FP + TN}$$

$$TPR = \frac{TP}{P} = \frac{TP}{FN + TP}$$

En contraste con FPR, TPR provee información útil sobre la proporción de ejemplos positivos identificados correctamente del total de positivos (P).

Tasa de verdaderos negativos o Especificidad: $1 - \frac{FP}{TN+FP}$

Las métricas de rendimiento de la precisión (Performance Metric Precision, PRE) y de la exhaustividad (Recall, REC) están relacionadas con esos índices TP y TN; de hecho, REC es sinónimo de TPR:

$$REC = TPR = \frac{TP}{P} = \frac{TP}{FN + TP}$$

En otras palabras, la exhaustividad cuantifica cuantos de los registros relevantes (los positivos) se capturan como tales (los verdaderos positivos). La precisión cuantifica cuántos de los registros pronosticados como relevantes (la suma de los verdaderos y falsos positivos) son realmente relevantes (los verdaderos positivos):

$$PRE = \frac{TP}{TP + FP}$$

Para equilibrar los aspectos positivos y negativos de la optimización de PRE y REC, se utiliza la media armónica de PRE y REC, llamada puntuación F1:

$$F1 = 2\frac{PRE\ X\ REC}{PRE + REC}$$

Por último, una medida que resume la matriz de confusión es MCC (Matthews Correlation Coefficient), se calcula como sigue:

$$MCC = \frac{TP\ X\ TN - FP\ X\ FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

A diferencia de PRE, REC y la puntuación F1, MCC oscila entre -1 y 1, y tiene en cuenta todos los elementos de la matriz de confusión; por ejemplo, la puntuación F1 no tiene en cuenta TN. Aunque los valores de MCC son más difíciles de interpretar que la puntuación F1, se considera una métrica superior (Raschka et al., 2023).

Representación de la característica operativa del receptor

"La figura de la característica operativa del receptor (Receiver Operating Characteristics, ROC) es una herramienta útil para seleccionar la técnica de clasificación en función de su rendimiento con respecto a FPR y TPR, que se calculan desplazando el umbral de decisión del clasificador. La diagonal de un gráfico ROC puede interpretarse como una suposición aleatoria, y las técnicas de clasificación que caen por debajo de la

diagonal se consideran peores que la suposición aleatoria. Un clasificador perfecto se situaría en la esquina superior izquierda del gráfico con valores de TPR de 1 y FPR de 0. Basándonos en la curva ROC, podemos calcular la llamada área ROC bajo la curva (ROC AUC) para determinar el rendimiento de las técnicas de clasificación".

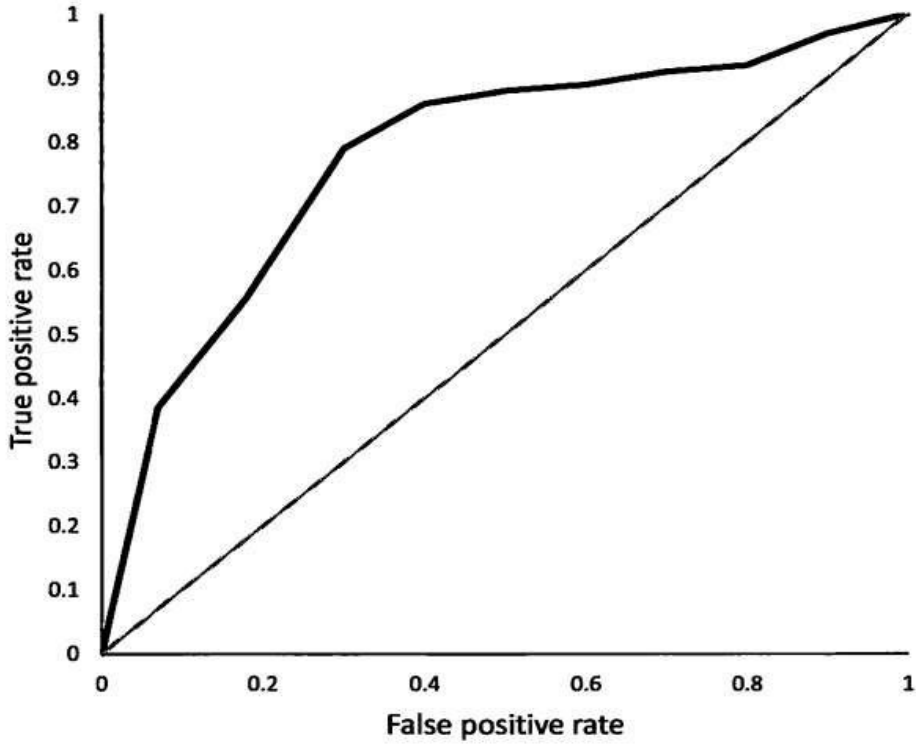


Figura 11. Característica operativa del receptor.

2.2.13 Importancia de las variables con la técnica Random Forest

Criterio de división

La técnica "Random Forest" selecciona la división con menor impureza en cada nodo utilizando la medida de impureza de Gini (Breiman, 2003). La impureza de Gini cuantifica la forma en que se distribuyen los identificadores de clase en un nodo determinado. El valor de la "impureza de Gini" oscila entre 0 y 1, siendo 0 la condición de que todos los elementos de un nodo pertenezcan a la misma clase. La medida de impureza de Gini se define formalmente del siguiente modo: Si $X = x_1, x_2,..., x_j$ en el nodo t representa el número de mujeres adolescentes en el nodo t, N denota el número de muestras, n_{ci} significa el recuento de muestras con valor x_i que pertenecen a la clase c, y a_i representa el recuento de muestras con valor x_i en el nodo t, entonces la "impureza de Gini" puede calcularse como sigue: (Breiman, 1984)

$$I\left(t_{x_i}\right) = 1 - \sum_{c=0}^{c} \left(\frac{n_{ci}}{a_i}\right)^2 \quad \text{(a)}$$

Suponiendo que sea la media ponderada de la medida de Gini entre los distintos valores de la variable X, el índice de Gini de una división puede calcularse de la siguiente manera:

$$Gini\,(t, X) = \sum_{i=1}^{j} \frac{a_i}{N} I\left(t_{x_i}\right) \quad \text{(b)}$$

El criterio de división vendrá determinado por cuál de las m variables tiene el menor valor de impureza de Gini cuando se calcula. El algoritmo "Random forest" construye el criterio de división para cada árbol utilizando un conjunto distinto de m variables.

Importancia de las variables

La "importancia de una variable" cuantifica el grado en que una determinada variable influye en el resultado de la clasificación. El algoritmo "Random Forest" utiliza cuatro métricas para evaluar la importancia de una variable: el índice de Gini, la precisión decreciente y la puntuación de importancia absoluta para la clase 1. Para determinar la importancia de una variable j, el árbol procesa muestras fuera de la bolsa (OOB) y documenta la precisión de las predicciones. Tras permutar los valores de la variable j en las muestras OOB, se determina de nuevo la precisión. Estos cálculos se ejecutan de forma secuencial a lo largo de cada árbol individual a medida que se construye el Bosque Aleatorio. Al promediar la disminución de la precisión de estas permutaciones en todos los árboles, se puede determinar la importancia de la variable j. Una disminución sustancial de la precisión de la predicción indica que la variable j es significativa. Una disminución sustancial en la precisión de la predicción indica una fuerte asociación entre la respuesta y la variable j (Hastie, 2009). Una vez determinada la importancia de todas las variables, el "bosque aleatorio" genera una lista ordenada de "importancia de las variables".

En un sentido formal, considere lo siguiente: β_t representa las muestras OOB para el árbol t, donde t, $t \in \{1,..., ntree\}$, y'^{t}_{i} denota la clase predicha para la instancia i antes de la permutación en el árbol t y $y'^{t}_{i,\alpha}$ significa la clase predicha para la instancia i después

de la permutación. La siguiente expresión define la "importancia variable" (variable independiente) de la variable j en el árbol t:

$$VI_j^t = \frac{\sum_{i=1}^{N} \beta_t I\left(y_i = y_t^{'t}\right)}{|\beta_t|} - \frac{\sum_{i=1}^{N} \beta_t I\left(y_i = y_{i,\sigma}^{'t}\right)}{|\beta_t|} \quad \text{(c)}$$

El “valor de importancia” sin procesar para la variable j se promedia sobre todos los árboles en el “random forest”.

En el “random forest”, la media del "valor de importancia" original de la variable j se calcula en todos los árboles.

$$VI_j = \frac{\sum_{t=1}^{ntree} VI_m^t}{ntree} \quad \text{(d)}$$

La "importancia variable" viene determinada por el Gini de Disminución Media “Mean Decrease Gini” (MDG), que se establece según el criterio de división de Gini. La MDG cuantifica la reducción de ΔI (ecuación a) asociada al proceso de separación. El valor del cambio en I (ecuación f) en el nodo t para un problema de dos clases se calcula restando la media ponderada de la medida de Gini (ecuación b) de la impureza de clase (ecuación e) (Bjoern, 2009; Mingers, 1989).

$$I(t) = 1 - \sum_{c=0}^{c} \left(\frac{n_i}{N}\right)^2 \quad \text{(e)}$$

$$\Delta I(t) = I(t) - Gini(t, X) \quad \text{(f)}$$

La reducción de la "impureza de Gini" se documenta para cada nodo t en todos los árboles (ntree) del "random forest" para cada variable. Posteriormente, la importancia de Gini (IG) se calcula de la siguiente manera: (Bjoern, 2009; Khalilia et al., 2011)

$$GI = \sum_{ntree} \sum_{t} \Delta I(t) \quad \text{(g)}$$

2.3 Definición de términos básicos

A continuación, se detalla un glosario de términos que serán utilizados en el presente trabajo.

Embarazo adolescente: se caracteriza por transcurrir en una mujer durante el periodo que va desde el inicio de su edad reproductiva hasta la conclusión de la adolescencia. “La Organización Mundial de la Salud define la adolescencia como el periodo comprendido entre los 10 y los 19 años de edad” (Unimedicos, 2019).

Modelos machine learning supervisados: son modelos predictivos; identifican el modelo óptimo aprendiendo de un conjunto de datos en el que los resultados ya se han clasificado como positivos o negativos. Utilizando el modelo identificado como el más adecuado, se predicen los resultados de un conjunto de datos adicional etiquetado, denominado conjunto de datos de prueba. En el contexto del machine learning, el conjunto de datos para una tarea determinada suele dividirse de la siguiente manera: un conjunto de entrenamiento, que se utiliza para entrenar el modelo; un conjunto de prueba, que se emplea para estimar los parámetros que deben incorporarse a los modelos de entrenamiento; y un conjunto de datos de validación, que se utiliza para verificar la similitud entre los resultados previstos y los reales. El 20% de la muestra se asigna aleatoriamente al conjunto de datos de prueba. En el caso del 80% restante, se emplea la validación cruzada k-fold como técnica para evaluar la capacidad de generalización del modelo a un conjunto de datos independiente. El conjunto de validación se forma a partir de uno de los k subconjuntos de tamaño aproximadamente igual entre los que se dividen los datos en este método. Como datos de entrenamiento se utilizan los k-1 subconjuntos restantes.

Algoritmo MissForest: representa un enfoque basado en árboles que se emplea para imputar los datos que faltan. Se trata de técnicas no paramétricas que permiten imputar valores de variables tanto cuantitativas como categóricas.

Según el algoritmo MissForest, introducido por Stkhoven y Bühlmann, los métodos paramétricos pueden no captar las intrincadas interrelaciones y las estructuras no lineales de las relaciones que existen en determinados grupos de datos. El número de iteraciones del algoritmo y la profundidad máxima de los bosques son dos ejemplos de parámetros.

Este método emplea la estructura del árbol para realizar estimaciones de los valores faltantes.

El algoritmo es el siguiente:

1. El grupo de datos original se utiliza para generar múltiples conjuntos de datos completos. En ausencia de valores en estas categorías, se añaden valores iniciales. La asignación inicial puede diferir entre iteraciones del algoritmo debido a la utilización de métodos bootstrap.
2. Para pronosticar los valores que faltan, se aplica a cada conjunto de datos un modelo de clasificación construido a partir de árboles de decisión.
3. Para imputar los valores ausentes, se utilizan las predicciones del paso anterior.
4. Proceda iterando a través de los pasos 2 y 3 hasta que se alcance la convergencia.

Una de las ventajas de este modelo es su capacidad para imputar simultáneamente múltiples variables, incluidas variables categóricas. Además, destaca la adaptabilidad y escalabilidad del algoritmo, ya que puede aplicarse tanto a conjuntos de datos pequeños como grandes manteniendo un alto nivel de eficiencia computacional (Stekhoven & Bühlmann, 2012).

AUC: ("area under the curve", área bajo de la curva) "para comparar los clasificadores, se calcula el área bajo la curva ROC, tomando un valor de 1 cuando es un clasificador perfecto, mientras que para un clasificador estrictamente aleatorio es 0,5".

Pérez (2015): "explica que el área bajo la curva de la diagonal principal, que corresponde a la peor prueba diagnóstica, es de 0,5. Además, se han establecido los siguientes intervalos para los valores AUC de la curva ROC:

- [0.5 - 0.6[: Modelo malo
- [0.6 - 0.75[: Modelo regular
- [0.75 - 0.9[: Modelo bueno
- [0.9 - 0.97[: Modelo muy bueno
- [0.97 – 1[: Modelo excelente

Se plantean las siguientes hipótesis:

H0: AUC ROC es igual a 0.5

H1: AUC ROC no es igual a 0.5

Un rechazo de Ho acompañado del p-valor indica que el modelo ajustado es adecuado" (Pérez, 2015).

2.4 Hipótesis y variables

Hipótesis

Los factores más importantes que predicen el embarazo adolescente a través de la mejor técnica del machine learning en el Perú, durante los años 2019 al 2020 son: el haber tenido una primera unión y si alguna vez usó cualquier método anticonceptivo.

Variables

Se definen las variables con respecto a las preguntas de la ENDES 2019-2020, así:

Tabla 1. *Operacionalización de variables*

Factores y variables	**Etiqueta**	**Tipo de variable**	**Valores finales (codificación)**
Variable dependiente:			
Embarazo adolescente	MA	Cualitativa	0: No 1: Si
Variables independientes:			
Variables Intermedias:			
Edad en la primera relación sexual	primera	Cuantitativa	9, 10, …
Ha tenido una primera unión	union	Cualitativa	0: No 1: Si
Ha estado casada o ha convivido	casada	Cualitativa	0: No 1: Si
Alguna vez usó cualquier método	anticoncep	Cualitativa	0: No 1: Si
Uso actual de métodos anticonceptivos	uso	Cualitativa	0: No 1: Si
Factores subyacentes:			
Factores socioculturales y ambientales:			
Etnia	etnia	Cualitativa	1: Castellano u otra lengua extranjera 2: Otra lengua nativa u originaria 3: Quechua
Área de residencia	area	Cualitativa	0: Rural 1: Urbana
Región de residencia	region	Cualitativa	1: Lima Metropolitana 2: Resto Costa 3: Selva 4: Sierra
Factores familiares:			
Número de miembros del hogar	miembros	Cuantitativa	0, 1, …
Índice de riqueza	riqueza	Cualitativa	1: Más rico 2: Rico 3: Medio 4: Pobre 5: Más pobre

Factores individuales:			
Migración	migracion	Cualitativa	0: No 1: Si
Conocimiento del ciclo ovulatorio	ciclo	Cualitativa	0: No 1: Si
Han escuchado de planificación familiar (pf) por medios de comunicación	planificacion	Cualitativa	0: No 1: Si
Número deseado de hijos	deseado	Cuantitativa	0, 1, …
Edad (en años cumplidos)	edad	Cuantitativa	12, …,19
Empleo en los últimos 12 meses	empleo	Cualitativa	0: No 1: Si
Educación (en años)	educacion	Cuantitativa	0, 1, …
Actualmente asiste a la escuela, colegio, instituto o universidad	asiste	Cualitativa	0: No 1: Si

Capítulo III

METODOLOGÍA

3.1 Tipo y Diseño de Investigación

Esta investigación es de tipo básica, ya que tiene como propósito producir conocimiento y teorías (Hernández – Sampieri & Mendoza, 2018).

Asimismo, es de diseño no experimental, de corte transversal, de alcance correlacional - causal. Al no manipular de forma deliberada las variables de estudio es no experimental. También, es correlacional–causal porque busca describir correlaciones entre variables o relaciones causales entre categorías o variables en uno o más grupos de personas, es transversal ya que los datos se recolectan en un tiempo único (en un solo momento). Este diseño nos brinda la oportunidad de predecir el comportamiento de una o más variables a partir de otras, una vez que se establece la causalidad (Hernández-Sampieri & Mendoza, 2018).

3.2 Unidad de Análisis

Una mujer adolescente de 12 a menos de 20 años de edad del Perú en los años 2019-2020.

3.3 Población de Estudio

Los datos utilizados en esta investigación se obtuvieron de la ENDES 2019-2020. Esta encuesta se lleva a cabo en más de noventa países con el fin de comprender las tendencias demográficas y sanitarias en los países en desarrollo.

La población objeto de estudio fueron las mujeres adolescentes peruanas de 12 a menos de 20 años de edad en los años 2019-2020.

3.4 Tamaño de Muestra

La muestra consistió en todas las mujeres adolescentes (16825) que tenían entre 12 y menos de 20 años en el momento de la entrevista (durante el año escolar 2019-2020) con la ENDES.

3.5 Selección de Muestra

Todas las mujeres de 12 a menos de 20 años de edad que residen en el Perú encuestadas por el INEI utilizando la ENDES 2019-2020.

3.6 Técnicas de Recolección de Datos

Se usó datos de la ENDES 2019-2020. Esta encuesta se realiza todos los años y entrevista a mujeres de entre 12 y 49 años. Incluye preguntas sobre las características sociodemográficas de la encuestada, su historial de matrimonio y convivencia, uso de anticonceptivos, servicios de planificación familiar, contactos con trabajadores de salud comunitarios, embarazo, parto, atención posnatal y nutrición infantil, preferencias de fertilidad de la mujer y otros problemas de salud. Este análisis incluye a las mujeres que informaron haber estado casadas alguna vez y tenían entre 12 y menos de 20 años de edad, para permitir el análisis del embarazo adolescente como se practica más recientemente. El estándar internacional para las estimaciones de prevalencia del embarazo adolescente se basa en la prevalencia de su ocurrencia entre las personas de 12 a menos de 20 años, porque esta edad ofrece tiempo suficiente para que se produzca un embarazo adolescente (INEI, 2020).

3.7 Análisis e Interpretación de la Información

Para el análisis de la información se utilizó lo recogido en el módulo de salud, vivienda y hogar con el cuestionario individual de la ENDES 2019-2020.

Nuestro resultado primario es el embarazo adolescente, según lo informado por la mujer. Se clasificó a aquellas que informaron haber contraído embarazo a una edad de 12 hasta antes de los 20 años como casos de embarazo adolescente.

Para este estudio se adoptó un alcance explicativo, con la intención de descubrir la importancia de los predictores en su asociación con el embarazo adolescente entre mujeres de 12 a menos de 20 años. Por tanto, las variables se seleccionaron de las secciones de la ENDES 2019-2020 de acuerdo al marco teórico propuesto por Di Cesare

e Rodríguez-Vignoli (2006), las que se incluyeron en las técnicas de Machine Learning. Sin embargo, para garantizar la interpretabilidad de los resultados, las variables de datos se preprocesaron antes de su aplicación.

Primero, con los datos de la encuesta ENDES se revisó cada variable en el conjunto de datos y a) se eliminó ciertas variables que no sean significativas para el análisis, por ejemplo, variables relacionadas con la fecha de la entrevista, la identificación del caso y el conglomerado ID, y b) se identificó las diferentes categorías que debe incluir cada variable. La utilización de las categorizaciones de variables empleadas en investigaciones anteriores facilitó la interpretación de los resultados de forma coherente con la bibliografía existente.

Una vez que se identificaron las categorías de las variables, se codificó las variables categóricas.

Segundo, se imputo los datos faltantes usando la función missForest de la librería y paquete missForest del software estadístico R.

Tercero, se realiza el análisis exploratorio de datos (EDA) univariante y bivariado con el software estadístico Python.

Cuarto, conversión de datos categóricos en números utilizando la codificación One Hot con el enfoque Label Encoding.

Quinto, partición de los datos en conjuntos de entrenamiento y prueba.

Sexto, escalado o normalización de los datos.

Sétimo, implementación de los nueve algoritmos utilizados para la presente tesis.

Todos los modelos se ejecutaron o entrenaron primero en el conjunto de datos de entrenamiento. Estimamos las métricas para la evaluación de la precisión del modelo y las tasas de error utilizando el conjunto de datos de prueba.

Comparamos nueve estrategias diferentes para proporcionar una mayor comprensión de las características que son significativas en los enfoques: Máquina de Soporte Vectorial, Regresión Logística Binaria, Árbol de decisión, AdaBoost, Gradient Boosting, XGBoost, ExtraTrees, Bagging y Random Forest. Estos modelos se seleccionaron en función de su popularidad constante y duradera en el machine learning para tareas de clasificación (Stock et al., 2012; Muthukrishnan & Rohini, 2016) y su uso

reconocido y recomendado en la salud pública y la investigación en ciencias del comportamiento (Mooney & Pejaver, 2018; McNeish, 2015; Seligman et al., 2018; Ghaoui et al., 2010). Al emplear nueve enfoques distintos de machine learning para analizar el mismo conjunto de datos, también es posible realizar una evaluación más exhaustiva de los resultados.

Para probar el rendimiento de nuestros modelos, se evaluó en busca de un indicador de precisión y error. Calculamos el AUC ROC, que se utiliza como indicador de precisión. La ROC es un gráfico de la tasa de verdaderos positivos de la prueba (eje Y) frente a la correspondiente tasa de falsos positivos (eje X); es decir, sensibilidad frente a especificidad. AUC ROC, se utiliza a menudo para resumir el rendimiento de prueba de los modelos. De manera simplista, AUC es la posibilidad de que una observación seleccionada al azar (es decir, un participante de la encuesta) que tuvo un embarazo adolescente se clasifique como esa (Hanley y McNeil, 1982).

Octavo, determinar la importancia de los predictores en el mejor modelo encontrado.

Tanto para el análisis de datos y el procesamiento de los mismos, se utilizó los softwares IBM SPSS versión 26, R con su interfaz R Studio y Python con su interfaz la aplicación Google Colaboratory con las bibliotecas necesarias (pandas, scipy, keras, numpy, sklearn, tensorflow) para desarrollar los algoritmos predictivos con el fin de identificar variables asociadas con el embarazo en adolescentes peruanas, así como también la mejor técnica machine learning que describa mejor el embarazo adolescente de acuerdo a sus métricas.

CAPITULO 4

RESULTADOS Y DISCUSIÓN

4.1 Resultados

Características de la muestra

Para mejorar la comprensión y familiarizarse con las adolescentes que componen nuestra base de datos, comenzamos proporcionando tablas descriptivas que abarcan atributos sociodemográficos y de salud de las 16825 adolescentes cuyos datos se recopilaron a través de la ENDES de 2019 a 2020.

En la Tabla 2 se observa, que el embarazo (mujeres alguna vez embarazadas entre los 12 a 19 años) ocurrió en 1830 (10.9%) de 16825 adolescentes.

Tabla 2. *Distribución de la variable embarazo adolescente*

Embarazo Adolescente	Frecuencia	Porcentaje
No	14995	89.1
Si	1830	10.9
Total	16825	100.0

Fuente: Elaboración propia.

La tabla 3, presenta las variables sociodemográficas y de salud de la muestra de adolescentes de 12 a 19 años de edad; la mayoría no han tenido una primera unión (90.6%), no ha estado casada o convivido (98.6%), no han usado alguna vez cualquier método (80.5%), no usan actualmente métodos anticonceptivos (89.6%), su lengua materna es castellano u otra lengua extranjera (87.9%), residen en el área urbana (63.6%), vive en la sierra (36.2%), viven en el resto de la Costa y en la Selva (26.6% y 26.4%, respectivamente)pertenece al índice de riqueza más pobre (34.9%), pertenece al índice de riqueza pobre (25.8%), no se han desplazado (72.7%), no tienen conocimiento del ciclo ovulatorio (87.0%), no han escuchado de planificación familiar por medios de comunicación (69.5%), no han tenido empleo en los últimos 12 meses (75.4%), y si asisten a la escuela, colegio, instituto o universidad (78.9%).

Por otro lado, el 87.1% de las adolescentes que han tenido una primera unión ya han tenido un embarazo en comparación con las que no han tenido una primera unión (3.0%), el 84.6% de las adolescentes que han estado casadas o convivido han tenido un embarazo en comparación con las que no han estado casadas o convivido (9.8%). En relación a las adolescentes que usaron alguna vez cualquier método anticonceptivo, el 51.1% ya han tenido un embarazo en comparación con el 1.1% en aquellas que no usaron; en cuanto a los que usan actualmente métodos anticonceptivos, el 61.3% de las adolescentes ya han tenido un embarazo en comparación con las que no usan actualmente (5.0%); un 16.6% de las adolescentes cuya lengua materna es otra lengua nativa u originaria ya han tenido un embarazo.

En cuanto al área de residencia, el 13.3% de las adolescentes que residían en el área rural ya han tenido un embarazo en comparación con las que residían en el área urbana (9.5%), en la región de residencia de la selva el 13.9% también ya han tenido un embarazo, seguido por un 10.7% en el resto de la Costa, un 9.7% en la Sierra y un 7.9% en Lima Metropolitana. Además, el 14.0% de las adolescentes provenientes de hogares más pobres ya han tenido un embarazo, en hogares pobres este fenómeno ocurre en 12.3% de las adolescentes, un menor porcentaje ocurre en hogares ricos (4.1% de las adolescentes). El 18.6% de las adolescentes que se han desplazado han tenido un embarazo en comparación con las que no se han desplazado (8.0%); el 13.4% de las adolescentes que si conocen el ciclo ovulatorio han tenido un embarazo en comparación con las que no tienen conocimiento (10.5%).

Asimismo, un 11.8% de las adolescentes que no han escuchado de planificación familiar por medios de comunicación han tenido un embarazo adolescente en comparación con la que han escuchado (9.4%), un 22.2% de las adolescentes que han tenido empleo en los últimos 12 meses ya han tenido un embarazo en comparación con las que no han tenido empleo (7.2%). Por último, un 41.7% de las que no asisten a la escuela, colegio, instituto o universidad ya han tenido un embarazo adolescente en comparación con las que si asisten (2.6%).

Tabla 3. *Características sociodemográficas y de salud de las adolescentes según embarazo. Perú: 2019 al 2020*

Variable	Adolescentes	Porcentaje	Embarazo adolescente	% de embarazos adolescentes
Ha tenido una primera unión				
No	15240	90.6	450	3.0%
Si	1585	9.4	1380	87.1%
Ha estado casada o ha convivido				
No	16584	98.6	1626	9.8%
Si	241	1.4	204	84.6%
Alguna vez usó cualquier método				
No	13538	80.5	151	1.1%
Si	3287	19.5	1679	51.1%
Uso actual de métodos anticonceptivos				
No	15075	89.6	758	5.0%
Si	1750	10.4	1072	61.3%
Etnia				
Castellano u otra lengua extranjera	14787	87.9	1598	10.8%
Otra lengua nativa u originaria	391	2.3	65	16.6%
Quechua	1647	9.8	167	10.1%
Area de residencia				
Rural	6124	36.4	814	13.3%
Urbana	10701	63.6	1016	9.5%
Región de residencia				
Lima metropolitana	1800	10.7	143	7.9%
Resto Costa	4481	26.6	479	10.7%
Selva	4445	26.4	617	13.9%
Sierra	6099	36.2	591	9.7%
Índice de riqueza				
Más rico	1414	8.4	58	4.1%
Rico	2184	13.0	142	6.5%
Medio	3019	17.9	275	9.1%
Pobre	4342	25.8	534	12.3%
Más pobre	5866	34.9	821	14.0%
Desplazamiento (Migración)				
No	12239	72.7	978	8.0%
Si	4586	27.3	852	18.6%
Conocimiento del ciclo ovulatorio				
No	14641	87.0	1538	10.5%
Si	2184	13.0	292	13.4%
Han escuchado de pf por medios de comunicación				
No	10567	69.5	1242	11.8%
Si	6258	13.0	588	9.4%
Empleo en los últimos 12 meses				
No	12693	75.4	914	7.2%
Si	4132	24.6	916	22.2%
Actualmente asiste a la escuela, colegio, inst. o univ.				
No	3554	21.1	1482	41.7%
Si	13271	78.9	348	2.6%

Fuente: Elaboración propia.

La Tabla 4, muestra las variables cuantitativas, donde se observa que la edad promedio en la primera relación sexual fue de 14.64 años y una mediana de 14 años, es

decir, el 50% de las adolescentes se encuentra alrededor de los 14 años, por lo que se supone que tiende a una distribución simétrica, el número promedio de miembros del hogar fue de 5.49, y su mediana es de 5 miembros con lo que se llega a suponer que tiende a una distribución asimétrica, el número promedio deseado de hijos fue de 1.9, la edad promedio fue de 15.22 años y su mediana es de 15 años, y el número promedio de años de educación en las adolescentes fue de 8.46 y su mediana es de 8 años. Las desviaciones estándar no son muy altas como se observa en la Tabla, por lo que se puede decir que no existe mucha dispersión.

Tabla 4. *Estadísticas descriptivas de las variables cuantitativas de las adolescentes. Años: 2019 al 2020*

Variables independientes	Media	Desv. Est.	Mínimo	Mediana	Máximo
Edad en la primera relación sexual	14.64	1.218	9	14	19
Número de miembros del hogar	5.49	2.082	1	5	24
Número deseado de hijos	1.9	0.624	0	2	8
Edad (En años cumplidos)	15.22	2.258	12	15	19
Educación (en años)	8.46	2.226	0	8	15

Fuente: Elaboración propia.

Análisis exploratorio de datos

Según la Figura 12, se observa que el embarazo adolescente presenta una mayor correlación con las variables: ha tenido una primera unión (r = 0.79), alguna vez usó cualquier método (r = 0.64), Uso actual de métodos anticonceptivos (r = 0.55) y edad en años cumplidos (r = 0.41). Asimismo, se observa una correlación fuerte positiva entre la edad en años cumplidos y edad en la primera relación sexual (r = 0.76), del mismo modo se observa una relación fuerte entre la edad en la primera relación sexual con los años de educación de la adolescente (r = 0.72). Por último, existe una correlación inversa entre el embarazo adolescente con el hecho de que actualmente asiste a la escuela, colegio, instituto o universidad (r = -0.51).

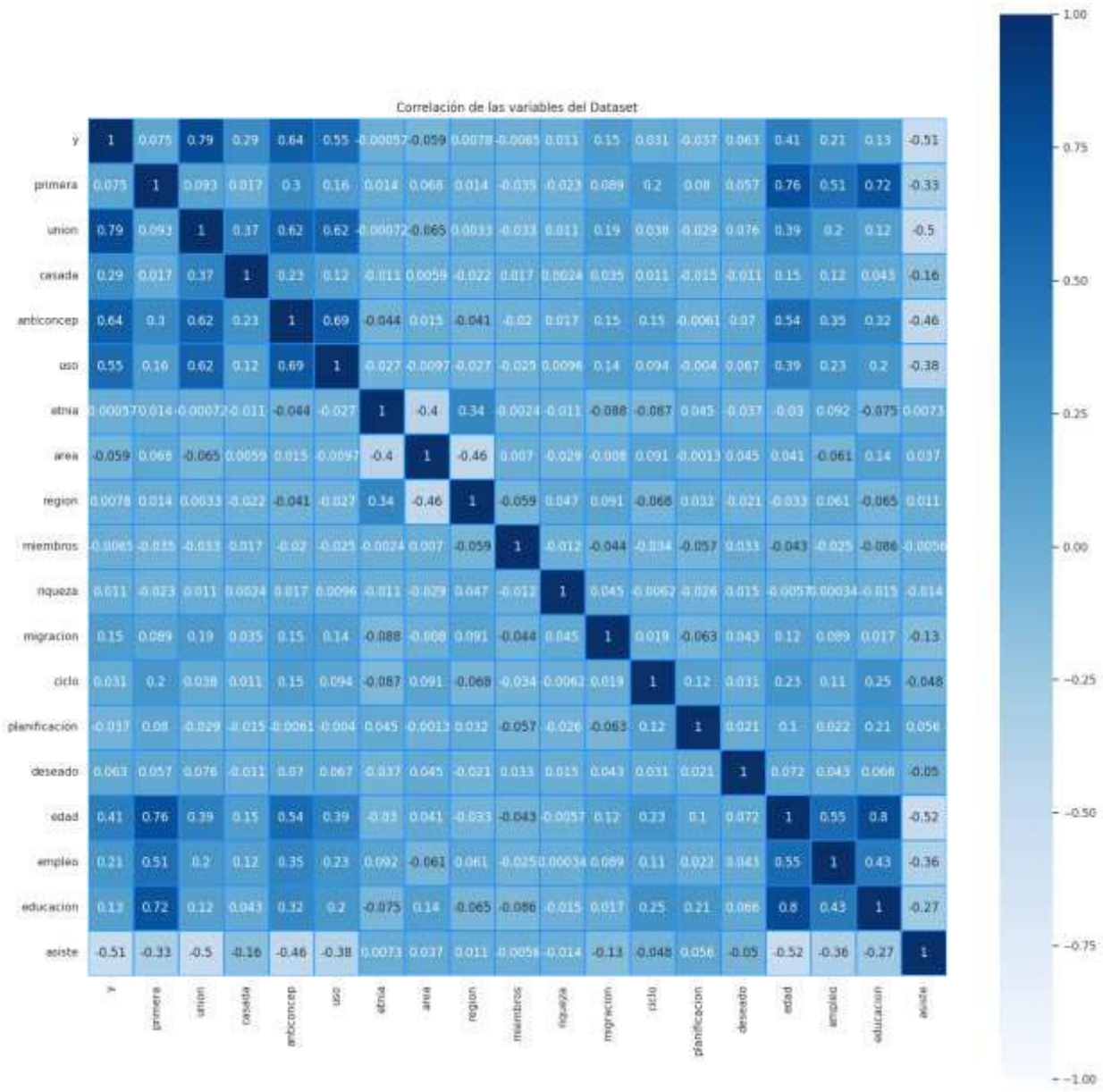


Figura 12. Correlación de las variables del estudio

El análisis con diagramas de cajas para identificar valores atípicos no presenta signos de alarma, realizándose también la normalización de los datos.

A continuación, aplicamos una partición estratificada de la muestra en función del objetivo para mantener la proporción de la variable dependiente tanto en los datos de entrenamiento como en los de evaluación. En concreto, los datos se dividieron del siguiente modo: El 80% se reservó para los datos de entrenamiento, y el 20% restante se utilizó para evaluar el rendimiento del modelo, la partición de la muestra se ejecutó utilizando la librería Sklearn Python, que se emplea sistemáticamente en la investigación para la construcción y comparación de modelos (consulte el anexo para ver el código).

Se observa el porcentaje de las categorías de la variable dependiente.

Tabla 5. *Tabla de contingencia del embarazo adolescente para los datos de entrenamiento*

Embarazo adolescente	Cantidad	% del total
0	11996	89.12%
1	1464	10.88%

"Fuente: Elaboración Propia".

Tabla 6. *Tabla de contingencia del embarazo adolescente para los datos de evaluación*

Embarazo adolescente	Cantidad	% del total
0	2999	89.12%
1	366	10.88%

"Fuente: Elaboración Propia".

La proporción original de la variable dependiente, se mantiene constante al comparar las Tablas 5 y 6, abarcando tanto los datos de entrenamiento (Train) como los de evaluación (para más detalles sobre el modelado algorítmico de datos, véase el anexo).

Ejecución de los modelos:

Con el objetivo de determinar y evaluar la predicción de los algoritmos de Máquina de Soporte Vectorial, Regresión Logística Binaria, Árbol de decisión, AdaBoost, Gradient Boosting, XGBoost, ExtraTrees, Bagging y Random Forest usando todas las variables y elegir el mejor resultado en la predicción del embarazo adolescente, se compararon las métricas de los algoritmos.

Comparación de modelos mediante métricas

1) Matriz de confusión de las clases estimadas y la variable dependiente

En la Tabla 7 se observa las matrices de confusión después de haber aplicado cada uno de los algoritmos en estudio, a partir de la cual se pueden calcular las diferentes métricas.

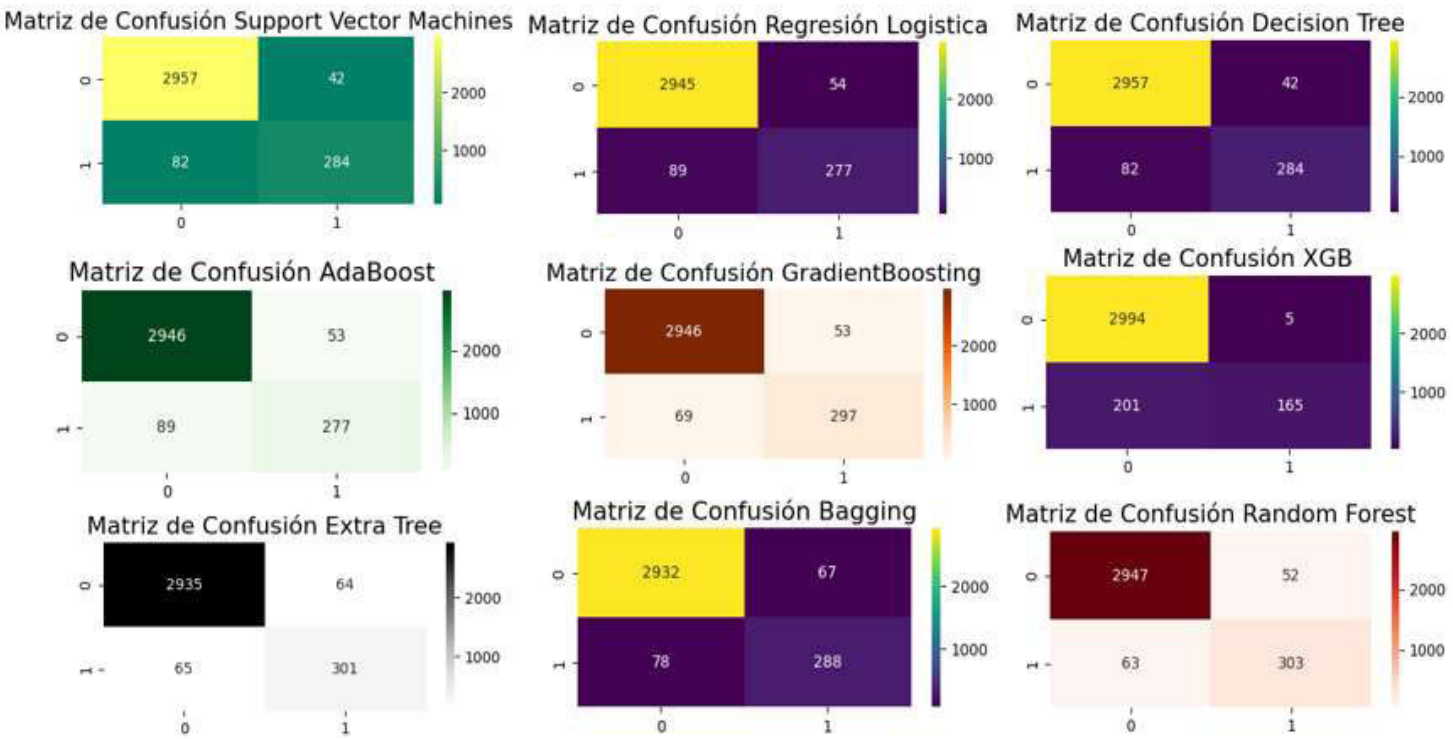


Figura 13. Matrices de confusión para todos los algoritmos.

2) Indicadores de comparación

En la tabla 7, muestran las métricas de las diferentes técnicas de clasificación obtenidos. Se encontró una precisión promedio ponderada (weighted avg) para los diferentes algoritmos con valores entre 0.94 y 0.97. Mostrando una mayor proporción promedio de predicciones positivas que fue realmente correcta para el algoritmo Random Forest (0.97), mientras que una menor proporción fue para el algoritmo XGBoost (0.94). El resto de algoritmos tuvo una precisión promedio de 0.96.

En cuanto a la métrica recall promedio ponderada (weighted avg) se observa que la proporción promedio de elementos positivos que se identificó correctamente fue mayor también con el algoritmo Random Forest (0.97) y la menor proporción fue para el algoritmo XGBoost (0.94). El resto de algoritmos tuvo un recall promedio de 0.96.

Por último, para la métrica f1-score promedio ponderada (weighted avg) que es una media armónica ponderada de la precisión y la recall se observa que quien presenta una proporción mayor también fue el algoritmo Random Forest (0.97) y la menor

proporción fue para el algoritmo XGBoost (0.93). Los demás algoritmos presentan un valor de 0.96.

Tabla 7. *Comparación de métricas de técnicas propuestas*

Algoritmos	Metricas de rendimiento															
	precisión				recall				f1-score				support			
	0	1	macro avg	weighted avg	0	1	macro avg	weighted avg	0	1	macro avg	weighted avg	0	1	macro avg	weighted avg
Máquina de Vectores Soporte	0.97	0.87	0.92	0.96	0.99	0.78	0.88	0.96	0.98	0.82	0.90	0.96	2999	366	3365	3365
Regresión Logística Binaria	0.97	0.84	0.90	0.96	0.98	0.76	0.87	0.96	0.98	0.79	0.89	0.96	2999	366	3365	3365
Árbol de decisión	0.97	0.87	0.92	0.96	0.99	0.78	0.88	0.96	0.98	0.82	0.90	0.96	2999	366	3365	3365
AdaBoost	0.97	0.84	0.91	0.96	0.98	0.76	0.87	0.96	0.98	0.80	0.89	0.96	2999	366	3365	3365
Gradient Boosting	0.98	0.85	0.91	0.96	0.98	0.81	0.90	0.96	0.98	0.83	0.90	0.96	2999	366	3365	3365
XGBoost	0.94	0.97	0.95	0.94	1.00	0.45	0.72	0.94	0.97	0.62	0.79	0.93	2999	366	3365	3365
ExtraTrees	0.98	0.82	0.90	0.96	0.98	0.82	0.90	0.96	0.98	0.82	0.90	0.96	2999	366	3365	3365
Bagging	0.97	0.81	0.89	0.96	0.98	0.79	0.88	0.96	0.98	0.80	0.89	0.96	2999	366	3365	3365
Random Forest	0.98	0.85	0.92	0.97	0.98	0.83	0.91	0.97	0.98	0.84	0.91	0.97	2999	366	3365	3365

Fuente: Elaboración propia.

3) Indicador Accuracy

El algoritmo con mejor accuracy (proporción de predicciones correctas realizadas por el modelo sobre todo tipo de predicciones realizadas) es el random forest (0.965825), seguido por el algoritmo Gradient Boosting (0.963744) y el algoritmo Decision Tree y Support Vector Machines (0.963155 para ambos).

Tabla 8. *Accuracy para los diferentes algoritmos de Machine Learning*

Modelo	**Accuracy**
Random Forest	0.965825
Gradient Boosting	0.963744
Decision Tree	0.963150
Support Vector Machines	0.963150
Extra tree	0.961664
AdaBoost	0.957801
Regresión logística	0.957504
Bagging	0.956909
XGB	0.938782

Fuente: Elaboración propia.

4) Curva ROC e Indicador AUC

Los 9 algoritmos propuestos obtuvieron un indicador AUC mayor de 0.90, donde según (Pérez, 2015), de acuerdo a los intervalos propuestos para medir la calidad de predicción del modelo a partir de la curva ROC resultaría que los 9 modelos se encuentran en una calidad de Muy Buena a Excelente.

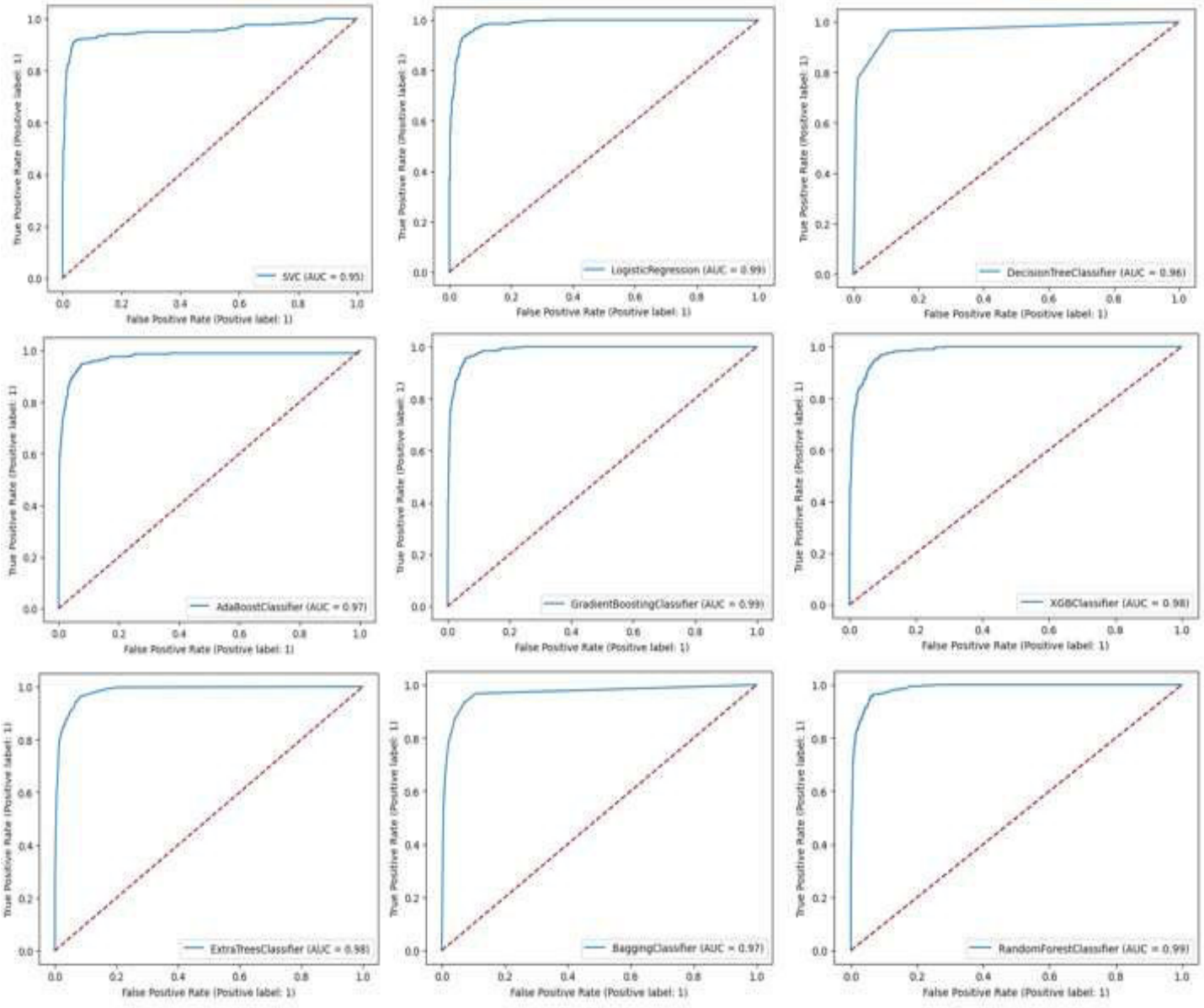


Figura 14. Curvas ROC para los algoritmos utilizados.

En cuanto el área bajo la curva, los algoritmos de Random Forest, Gradient Boosting y Regresión logística presentaron la mejor métrica, AUC de 0.99 para cada uno. Mientras que la menor AUC fue para el algoritmo Máquina de Vectores de soporte (0.95), para el resto de algoritmos su AUC se encuentra entre 0.96 y 0.98.

Tabla 9. *Comparación de los algoritmos*

Modelo	AUC
Random Forest	0.99
Gradient Boosting	0.99
Regresión logistica	0.99
Extra tree	0.98
XGB	0.98
AdaBoost	0.97
Bagging	0.97
Decision Tree	0.96
Support Vector Machines	0.95

Fuente: Elaboración propia.

5) Importancia de variables para la mejor técnica

Una de las capacidades de la técnica random forest es el calcular la importancia de las variables.

Tabla 10. *Importancia de las variables con la técnica Random Forest*

Variable	Descripción	Importancia (%)
union	Ha tenido una primera unión	30.9002%
anticoncep	Alguna vez usó cualquier método	15.2260%
asiste	Actualmente asiste a la escuela, colegio, instituto o universidad	7.4697%
uso	Uso actual de métodos anticonceptivos	7.4152%
primera	Edad en la primera relación sexual	6.9659%
edad	Edad (en años cumplidos)	5.7233%
miembros	Número de miembros del hogar	5.5032%
educacion	Educación (en años)	4.4718%
region	Región de residencia	2.6099%
riqueza	Índice de riqueza	2.5800%
deseado	Número deseado de hijos	2.2696%
empleo	Empleo en los últimos 12 meses	1.6407%
casada	Ha estado casada o ha convivido	1.6308%
migracion	Migración	1.3363%
planificacion	Han escuchado de planificación familiar (pf) por medios de comunicación	1.3359%
area	Área de residencia	1.0519%
ciclo	Conocimiento del ciclo ovulatorio	1.0066%
etnia	Etnia	0.8630%

De acuerdo a la Tabla 11 se observa que los 3 predictores más influyentes en orden de importancia para predecir el embarazo adolescente, acorde a los datos de entrenamiento son: ha tenido una primera unión (30.90%), seguido por Alguna vez usó

cualquier método anticonceptivo (15,23%) y Actualmente asiste a la escuela, colegio, instituto o universidad (7,47%).

Por otro lado, los 3 predictores que menos influyen en el embarazo adolescente son: Etnia (0,86%), conocimiento del ciclo ovulatorio (1,01%) y área de residencia (1,05%).

4.2 Discusión

Esta tesis planteó como objetivo determinar los factores más importantes que predicen el embarazo adolescente a través de la mejor técnica del machine learning en el Perú, durante los años 2019 al 2020, demostrándose un enfoque de la utilidad de técnicas de machine learning para clasificación y predecir resultados del embarazo adolescente. Debido a la edad relativamente joven en que esta posibilidad posterga a las adolescentes en la parte académica y laboral, hay un genuino interés en predecir el embarazo adolescente.

Se encontró que las variables más importantes fueron: Ha tenido una primera unión, seguido por Alguna vez usó cualquier método anticonceptivo y Actualmente asiste a la escuela, colegio, instituto o universidad, estos resultados son muy similares al estudio desarrollado por Quezada et al (2020) en Colombia quienes encuentran que las variables más influyentes son: Actualmente tiene pareja sexual, los métodos anticonceptivos han sido prescritos por una entidad promotora de salud (EPS), el individuo está matriculado en la escuela, inició la actividad sexual al menos a los quince años de edad, planifica sus relaciones y forma de educación, pero utilizando el algoritmo J48 del Machine Learning. Asimismo, Rosales (2019) también encontró que la variable más importante es la nupcialidad o unión temprana utilizando el modelo de fecundidad planteado por John Bongaarts. Además, Pinzón-Rondón et al. (2018) al utilizar la regresión logística encuentra que dentro de los principales factores para el embarazo adolescente se encuentran: convivir en pareja y asistir al colegio. Adicional a ello, estos resultados también corroboran lo encontrado por Binstock (2017) utilizando modelos de historia de eventos discretos, encuentra que la edad de inicio sexualmente y si se toman precauciones para prevenir un embarazo y la importancia que tiene la escuela en proveer información. Otro estudio como el de Avalos (2018) al utilizar la prueba chi cuadrado también encuentra que las relaciones sexuales inicien antes de los 16 años es importante para el embarazo adolescente se produzca. Finalmente, D'Añari (2019) con la aplicación de la

prueba Chi cuadrado encuentra que existe asociación estadística entre la edad de inicio de relaciones sexuales y el nivel de instrucción, con el embarazo adolescente.

Aunque existen escasas fuentes teóricas en el Perú en relación al empleo de las técnicas de machine learning en el estudio del embarazo adolescente para poder probar los resultados más ampliamente, siendo esta una limitación. Esta tesis demuestra que con la aplicación del machine learning se obtiene resultados que pueden utilizarse en el área de la salud y social. Por lo cual, en la presente investigación se aplicó nueve técnicas de machine learning (Regresión Logística Binaria, Máquina de Soporte Vectorial, Árbol de decisión, AdaBoost, Gradient Boosting, XGBoost, ExtraTrees, Bagging y Random Forest), encontrándose que la mejor técnica para predecir el embarazo adolescente fue el Random Forest con mejores métricas para la predicción (accuracy de 96.58% y área bajo la curva de 99%), considerándose como una predicción excelente, estos resultados se asemejan con lo encontrado por Rawat, et al (2021) quienes encuentran que la mejor técnica de Machine Learning a partir de datos de seguros médicos y de salud para la presentación y aceptación de reclamos fue Random Forest, no obstante, este resultado diverge de la investigación realizada por Oermann et al. (2016), que identificaron el modelo de regresión logística como el más eficaz con un accuracy del 74% para predecir la mortalidad de los pacientes sometidos a radiocirugía por malformaciones arteriovenosas cerebrales. Además, los hallazgos de Asadi et al. (2016) contradicen este resultado, ya que determinaron que las redes neuronales eran el modelo más eficaz (97,5% ± 2,5%) - técnica que no se ha tenido en cuenta para el desarrollo de la presente investigación -, además, estos resultados no se asemejan con la investigación realizada por Raj et al. (2020), que determinaron que el modelo más eficaz era el basado en redes neuronales (AUC=0,90) - técnica que no se ha tenido en cuenta para el desarrollo de la presente tesis -, sin embargo, cabe señalar que este estudio se centró en su predicción del matrimonio infantil en la India.

Una de las limitaciones de la presente tesis es el recojo de la información que realiza el INEI con su encuesta ENDES, para la selección de los individuos que pertenecen a la muestra no se tiene en cuenta la variable dependiente embarazo adolescente, como se sabe esta encuesta se desarrolla con la finalidad de calcular indicadores de desnutrición en la niñez y mortalidad en niños. Sin embargo, este estudio sería un punto de partida para la aplicación de las diferentes técnicas de machine learning para predecir diversas variables necesarias para una mejor gestión pública, donde se

requieren grandes bases de datos como son las que desarrolla el INEI que son accesibles para el público en general.

Se establece un nuevo paradigma de investigación con la aplicación de técnicas de machine learning a los datos sociales y sanitarios mediante el preprocesamiento de datos, la selección de variables que faciliten la generación de percepciones sociales y de acciones y decisiones que puedan aplicarse en la práctica social.

CONCLUSIONES

1. Los tres factores más importantes que ayudan a predecir el embarazo adolescente en un mayor porcentaje a partir de la técnica del Random Forest son: las variables intermedias (ha tenido una primera unión (30.90%), Alguna vez usó cualquier método anticonceptivo (15,23%)) y el factor subyacente (Actualmente asiste a la escuela, colegio, instituto o universidad (7,47%)).

2. El análisis descriptivo muestra altos porcentajes de embarazos adolescentes en aquellas que han tenido una primera unión (87,1%), han estado casadas o han convivido (84,6%), uso actual de métodos anticonceptivos (61,3%), alguna vez usó cualquier método (51,1%) y actualmente no asiste a la escuela, colegio, instituto o universidad (41,7%).

3. La métrica Precisión promedio ponderada (weighted avg) para los diferentes algoritmos presenta valores entre 0.94 y 0.97. Siendo mayor la proporción promedio de predicciones positivas que fue realmente correcta para el algoritmo Random Forest (0.97), mientras que una menor proporción fue para el algoritmo XGBoost (0.94). El resto de algoritmos tuvo una precisión promedio de 0.96. En el caso de la métrica Recall promedio ponderada (weighted avg) la proporción promedio de elementos positivos que se identificó correctamente fue mayor también con el algoritmo Random Forest (0.97) y la menor proporción fue para el algoritmo XGBoost (0.94). El resto de algoritmos tuvo un Recall promedio de 0.96. En cuanto el área bajo la curva, los algoritmos de Random Forest, Gradient Boosting y Regresión logística presentaron la mejor métrica (0.99 para cada uno), indicando que es una predicción Excelente para los embarazos adolescentes.

4. La mejor técnica de machine learning que determina los factores que influyen en el embarazo adolescente del Perú, durante los años 2019 al 2020 fue Random Forest ya que presenta un mayor Accuracy (0.965825).

RECOMENDACIONES

1. La incidencia de embarazos en adolescentes va en aumento, por lo que es necesario que los científicos realicen investigaciones que incorporen variables predictivas adicionales para facilitar las comparaciones y obtener resultados más fiables.
2. Se recomienda a los futuros tesistas, realizar un análisis exhaustivo de la optimización de hyperparámetros con la técnica del GridSearch en la mejor técnica de Machine Learning, teniendo en consideración que requiere un alto desgaste computacional.
3. En la actualidad, ciertos modelos de Machine Learning se consideran "cajas negras". Para optimizar con precisión las decisiones sanitarias, sería fundamental que los investigadores lleven a cabo estudios que utilicen técnicas de explicabilidad conductual e interpretación de algoritmos de Machine Learning con el fin de identificar los predictores más significativos del embarazo adolescente.
4. Se recomienda a los investigadores, desarrollar estudios que comparen diversas técnicas de balanceo con el fin de evaluar su potencial para predecir el embarazo en la adolescencia.
5. Se recomienda a los futuros investigadores, utilizar algoritmos de Deep Learning con redes neuronales que son técnicas de aprendizaje más profundas para predecir el embarazo adolescente, teniendo en cuenta el número de capas, las epochs, los nodos y todos sus demás hyperparámetros.

REFERENCIAS BIBLIOGRÁFICAS

Albornoz-Arias, N., Arenas-Villamizar, V. V., Martínez-Santana, M.-C., Carreño Paredes, M. T., & Sepúlveda-Aravena, J. (2019). *Factores socioecológicos para la intervención en embarazo de adolescentes en el estado Táchira, Venezuela*.

Albornoz-Arias, N., Arenas-Villamizar, V. V., Martinez-Santana, M.-C., Mora-Wilches, K., Fernández-Delgado, M.-K., Hernández Lalinde, J. D., Martínez-De-Biava, Y.-D.-C., & Sepulveda-Aravena, J. (2018). *Embarazo adolescente en el departamento Norte de Santander, Colombia. una mirada desde factores psicosociales para la intervención terapéutica*.

Arriaza, A. (2023). *Tasa de fecundidad de mujeres y adolescentes en Guatemala, año 2018: diferencias asociadas al contexto económico*. Universidad Rafael Landívar y UNFPA. https://guatemala.unfpa.org/sites/default/files/pub-pdf/fecundidad_completo.pdf

Asadi, H., Kok, F., Looby, S., Brennan, P., O'Hare, A., & Thornton, J. (2016). *Outcomes and Complications following Endovascular Treatment of Brain Arteriovenous Malformations-A Prognostication Attempt using Artficial Intelligence.* World Neurosurgery.

Avalos, L. E. P. (2018). *Factores asociados al embarazo en adolescentes atendidas en el centro de salud i-4 bellavista sullana 2017*. Tesis. https://repositorio.ucv.edu.pe/bitstream/handle/20.500.12692/26026/Avalos_VLEP.pdf?sequence=1&isAllowed=y

Azevedo, W. F. d., Diniz, M. B., Fonseca, E. S. V. B. d., Azevedo, L. M. R. d., & Evangelista, C. B. (2015). *Complications in adolescent pregnancy: systematic review of the literature*. Einstein (São Paulo), 13, 618–626.

Bellinger, C., Jabbar, M. S. M., Zaïane, O. & Osornio-Vargas, A. (2017). *A systematic review of data mining and machine learning for air pollution epidemiology*. BMC public health, 17(1), 1–19.

Binstock, G. (2017). *Maternidad adolescente en argentina: Factores asociados en contextos sociales vulnerables*. Anais, 1–22.

Bjoern, M, A. (2009). *Comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinformatics.* [PMC free article] [PubMed]

Bongaarts, J. (1978). *A framework for analyzing the proximate determinants of fertility*. Population and development review, 105-132.

Bratic, B., Kurbalija, V., Ivanovic, M., Oder, I., & Bosnic, Z. (2018). *Machine learning for predicting cognitive diseases: methods, data sources and risk factors*. Journal of medical systems, 42(12), 1–15.

Breiman, L. (1984). *Classification and regression trees*. Vol. 358. Wadsworth. Inc., Belmont, CA. [Google Scholar]

Breiman, L. (2003). *Manual-Setting Up, Using, and Understanding Random Forests V4. 0.* ftp://ftpstat.berkeley.edu/pub/users/breiman

Castañeda, J. y Santa-Cruz-Espinoza, H. (2021). *Factores de riesgo asociados al embarazo en adolescentes*. Enfermería Global vol. 20 N° 62 Murcia. https://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S1695-61412021000200004

D'Añari Cabrera, J. R. (2019). *Factores biosociodemográficos asociados al embarazo precoz en adolescentes gestantes atendidas en el HRHDE*, abril–mayo 2019.

Daoud, A., Kim, R. & Subramanian, S. (2019). *Predicting women's height from their socioeconomic status: A machine learning approach*. Social Science & Medicine, 238, 112486.

Davis, K. & Blake, J. (1956). *Social structure and fertility: An analytic framework*. Economic development and cultural change, 4(3), 211–235.

DeGregory, K., Kuiper, P., De Silvio, T., Pleuss, J., Miller, R., Roginski, J., Fisher, C., Harness, D., Viswanath, S. & Heymsfield, S. (2018). *A review of machine learning in obesity*. Obesity reviews, 19(5), 668–685. 35 36

Di Cesare, M. & Rodríguez Vignoli, J. (2006). *Análisis micro de los determinantes de la fecundidad adolescente en Brasil y Colombia*. Papeles de población, 12(48), 107–140.

García, A. (2005). *Métodos Avanzados de estadística aplicada*. Técnicas avanzadas. UNE. Madrid.

Glaser, B. G. & Strauss, A. L. (2017). *Discovery of grounded theory: Strategies for qualitative research*. Routledge.

Great-Learning-Team (2020). *A complete understanding of lasso regression*.

Hastie, T. (2009). *The elements of statistical learning: data mining, inference and prediction.* pp. 605–622.

Hernández – Sampieri, R. & Mendoza, C. (2018). *Metodología de la investigación. Las rutas cuantitativa, cualitativa y mixta*. Ciudad de México, México: Editorial Mc Graw Hill Education, ISBN: 978-1-4562-6096-5, 714 p.

Hilbe, J. (2009). *Logistic Regression Models*. Jet Propulsion Laboratory. California Institute of Technology and Arizona State University. Tempe, U. S. A. Chapman & Hall/CRC. Taylor & Francis Group.

INEI (2019). *Perú - Indicadores de Violencia Familiar y Sexual*, 2012-2019.

INEI (2021). *Encuesta de demografía y salud familiar 2020*.

Jeha, D., Usta, I., Ghulmiyyah, L. & Nassar, A. (2015). *A review of the risks and consequences of adolescent pregnancy*. Journal of neonatal-perinatal medicine, 8(1), 1–8.

Kaplanoglu, M., Bülbül, M., Konca, C., Kaplanoglu, D., Tabak, M. S., & Ata, B. (2015). *Gynecologic age is an important risk factor for obstetric and perinatal outcomes in adolescent pregnancies.* Women and Birth, 28(4), e119–e123.

Khalilia, M., Chakraborty, S. & Popescu, M. (2011). *Predicting disease risks from highly imbalanced data using random forest.* BMC Medical Informatics and Decision Making 2011, 11:51. doi: 10.1186/1472-6947-11-51. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3163175/

LeCun, Y., Bengio, Y., & Hinton, G. (2015). *Deep learning*. Nature, 521(7553), 436–444.

Loredo-Abdalá, A., Vargas-Campuzano, E., Casas-Muñoz, A., González-Corona, J., & De Jesús Gutiérrez-Leyva, C. (2017). *Embarazo adolescente: sus causas y repercusiones en la diada*. Revista Médica del Instituto Mexicano del Seguro Social, 55(2), 223–229.

Mingers, J. (1989). *An empirical comparison of selection measures for decision-tree induction. Machine learning*. 1989;3(4):319–342. [Google Scholar]

Mohammed, J. & Wagner, M. (2020). *Data Mining and Machine Learning. Fundamental concepts and algorithms.* Second edition. Cambridge University press. United Kingdom. DOI: 10.1017/9781108564175.

Mooney, S. J. & Pejaver, V. (2018). *Big data in public health: terminology, machine learning, and privacy*. Annual review of public health, 39, 95–112.

Oermann, E., Rubinsteyn, A., Ding, D., Mascitelli, J., Starke, R., Bederson, J., & Kondziolka, D. (2016). *Using a Machine Learning Approach to Predict Outcomes after Radiosurgery for Central Arteriovenous Malformations*. Scientific reports, 6.

Onofre, E. V. & Vargas, J. M. R. (2016). *Factores determinantes de la fecundidad en adolescentes atendidas en el hospital "Carlos Monge Medrano" de Juliaca*. Revista Científica Investigación Andina, 15(1), 99–106.

Pinzón-Rondón, Á. M., Ruiz-Sternberg, Á. M., Aguilera-Otalvaro, P. A., & Abril-Basto, P. D. (2018). *Factores asociados al inicio de vida sexual y al embarazo adolescente en Colombia. Estudio de corte transversal.* Revista chilena de obstetricia y ginecología, 83(5), 487–499. https://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0717-75262018000500487#B4

Quezada, M. A., Tobón-Rivera, A. & Castrillón-Gómez, O. D. (2020). *Minería de datos: una aplicación para determinar cuáles factores socio-económicos influyen en el embarazo adolescente*. Información tecnológica, 31(6), 53–60.

Raj, A., Dehingia, N., Singh, A., McDougal, L. & McAuley, J. (2020). *Application of machine learning to understand child marriage in India*. SSM - Salud de la población 12 (2020) 100687. https://www.sciencedirect.com/science/article/pii/S2352827320303244

Raj, A., Saggurti, N., Balaiah, D., & Silverman, J. G. (2009). *Prevalence of child marriage and its effect on fertility and fertility-control outcomes of young women in india: a cross-sectional, observational study*. The lancet, 373(9678), 1883–1889.

Rawat, S., Rawat, A, Kumar, D. & Sabitha A. (2021). *Application of machine learning and data visualization techniques for decision support in the insurance sector*. International Journal of Information Management Data Insights. Volume 1, Issue 2, November,100012. https://www.sciencedirect.com/science/article/pii/S2667096821000057

Raschka, S., Liu, Y. & Mirjalili, V. (2023). *Machine Learning con PyTorch y Scikit-Learn. Como desarrollar modelos de Machine Learning y Deep Learning con Python*. Editorial Marcombo. Primera edición.

Rodríguez Vignoli, J. (2014). *Fecundidad adolescente en américa latina: una actualización*.

Rosales, J. Y. (2019). *Determinantes próximos de la fecundidad adolescente en honduras periodo 2011-2012*. Población y Desarrollo-Argonautas y Caminantes, 15, 88–97. DOI: https://doi.org/10.5377/pdac.v15i0.8118

Seligman, B., Tuljapurkar, S., & Rehkopf, D. (2018). *Machine learning approaches to the social determinants of health in the health and retirement study*. SSM-population health, 4, 95–99.

Sánchez, H. & Reyes, C. (2002). *Metodología y diseños en la investigación científica*.

Thankachan, K. (2022). *What? When? How?: ExtraTrees Classifier*. https://towardsdatascience.com/what-when-how-extratrees-classifier-c939f905851c

Unimédicos (2019). *Embarazo en adolescente* (OMS). https://www.unimedicos.com/portal/articulo/embarazo-en-adolescentes-(oms)

Stekhoven, D.J. & Bühlmann, P. (2012). *Missforest – non-parametric missing value imputation for mixed-type data*, Bioinformatics, vol. 28, no. 1, pp. 112–118.

Véliz, C. (2018). *Aprendizaje automático. Análisis para la minería de datos y big data*. Departamento académico de ciencias. PUCP.

Yakubu, I., & Salisu, W. J. (2018). *Determinants of adolescent pregnancy in subSaharan Africa: a systematic review*. Reproductive Health, 15(1), 1-11.

Williamson, N. et al. (2013). *Estado de la población mundial 2013: Maternidad en la niñez. enfrentar el embarazo en adolescentes.* Technical report, Fondo de Población de las Naciones Unidas, New York (EUA).

Wilmott, P. (2019). *Machine Learning. An applied mathematics introduction*. First edition. Panda Ohana Publishing.

ANEXOS

Matriz de Consistencia

Problema	**Objetivos**	**Hipótesis**	**Variables**	**Técnica de recolección de datos**
Formulación del Problema: ¿Cuáles son los factores más importantes que predicen el embarazo adolescente a través de la mejor técnica del machine learning en el Perú, durante los años 2019 al 2020?	**Objetivo general:** Determinar los factores más importantes que predicen el embarazo adolescente a través de la mejor técnica del machine learning en el Perú, durante los años 2019 al 2020.	**Hipótesis:** Los factores más importantes que predicen el embarazo adolescente a través de la mejor técnica del machine learning en el Perú, durante los años 2019 al 2020 son: el haber tenido una primera unión y si alguna vez usó cualquier método anticonceptivo.	**Variable dependiente** Embarazo adolescente **Variables independientes** **Variables Intermedias** Edad en la primera relación sexual Ha tenido una primera unión Ha estado casada o ha convivido Alguna vez usó cualquier método Uso actual de métodos anticonceptivos **Factores subyacentes** **Factores socioculturales y ambientales** Etnia Área de residencia Región de residencia **Factores familiares** Número de miembros del hogar Índice de riqueza **Factores individuales** Migración Conocimiento del ciclo ovulatorio Han escuchado de PF por medios de comunicación Número deseados de hijos Edad (en años cumplidos) Vive con pareja Empleo en los últimos 12 meses Educación (en años) Actualmente asiste a la escuela, colegio, instituto o universidad	Encuesta
	Objetivos específicos: • Describir los factores intermedios y subyacentes según embarazo adolescente. • Calcular y comparar las métricas que permiten predecir el embarazo adolescente en el Perú a partir de las matrices de confusión obtenidas con las técnicas de machine learning. • Establecer la mejor técnica de machine learning que determina los factores que inciden en el embarazo adolescente del Perú, durante los años 2019 al 2020.			