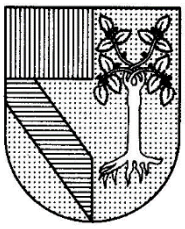


UNIVERSIDAD PANAMERICANA
FACULTAD DE INGENIERÍA

Con estudios incorporados a la
Secretaría de Educación Pública

“Metodología para determinar los factores de riesgo asociados con enfermedades complejas: Degeneración Macular Relacionada con la Edad y Preeclampsia.”

TESIS

QUE PARA OBTENER EL GRADO DE

DOCTOR EN INGENIERÍA

P R E S E N T A

ANTONIETA TEODORA MARTÍNEZ VELASCO

DIRECTORES:

DRA. MARÍA DE LOURDES GUADALUPE MARTÍNEZ VILLASEÑOR

DR. FRANCISCO JAVIER ESTRADA MENA

CDMX **2021**

AGRADECIMIENTOS

A mi directora de tesis, Dra. Lourdes Martínez Villaseñor.

A mi co-director de tesis, Dr. Javier Estrada Mena.

A la Facultad de Ingeniería de la Universidad Panamericana por brindarme los recursos y herramientas para llevar a cabo este trabajo.

A mi familia, por apoyarme en todo momento. En especial a Mario, quien siempre ha estado a mi lado como soporte e invaluable compañía.

A Andrea por ser una motivación para seguir.

A mis padres, quienes siempre me infundieron inspiración y confianza para alcanzar mis metas.

A Dios, por permitirme llegar a este punto.

Resumen

El incremento en la aplicación de la inteligencia artificial en la creación de sistemas de soporte de decisiones a escala está transformando también el futuro del cuidado de la salud. La inteligencia artificial se ha utilizado para implementar sistemas de diagnóstico y pronóstico de enfermedades, optimización del tratamiento y predicción del resultado, desarrollo de fármacos y para lidiar con problemas de salud pública (Noorbakhsh-Sabet, N., Zand, R., Zhang, Y., & Abedi, 2019).

La medicina personalizada hace énfasis en la prevención, el diagnóstico temprano y el tratamiento de las enfermedades complejas cuya incidencia se ha incrementado en los últimos años (Peek et al., 2015). Es necesario contar con métodos de diagnóstico rápidos y confiables que determinen los factores de riesgo asociados a las enfermedades que se estudien. Sin embargo, muchas de las tareas necesarias en el área de salud como la predicción de riesgos clínicos, el diagnóstico y la terapéutica son más retadores para las aplicaciones de inteligencia artificial (Bica, I., Alaa, A. M., Lambert, C., & van der Schaar, 2020; Maddox, T. M., Rumsfeld, J. S., & Payne, 2019).

Los datos provenientes de los pacientes se pueden obtener de los registros médicos; estos generalmente son colecciones complejas de datos, con muchas variables almacenadas, interacciones entre ellos y frecuentemente con un tiempo estricto de la validez. Adicionalmente, los datos de los pacientes según todos los factores de riesgo suelen no ser perfectos, ni completos (Singh Malik et al., 2007) y están generalmente desbalanceados. La determinación de los factores de riesgo es un reto importante debido a la gran cantidad de datos que actualmente se generan a partir de los estudios genéticos y datos clínicos obtenidos en la consulta médica de algunos hospitales.

Adicionalmente, con el fin de ofrecer una herramienta para apoyar a los expertos en el ámbito médico en la toma de decisiones, es necesario presentar los resultados obtenidos de manera que se dé transparencia acerca de los mecanismos subyacentes de los sistemas que se han utilizado para analizar los datos para generar conocimiento (Doran et al., 2018).

Con el fin de atender estos retos, en este trabajo se presenta una metodología para determinar los factores de riesgo asociados a enfermedades complejas mediante el enfoque

de aprendizaje automático. La metodología se probó en dos escenarios de aplicación: Degeneración Macular Relacionada con la Edad (DMRE) y Preeclampsia (PE). Es importante notar que la metodología incluye desde la generación de datos hasta la entrega un sistema de toma de decisiones interpretable.

Abstract

The increasing use of artificial intelligence in developing decision support systems at scale is also transforming the future of healthcare. Artificial intelligence has been used to implement disease diagnosis and prognosis systems, treatment optimization and outcome prediction, drug development, dealing with public health problems (Noorbakhsh-Sabet et al., 2019).

Personalized medicine emphasizes prevention, early diagnosis, and treatment of complex diseases whose incidence has increased in recent years (Peek et al., 2015). Therefore, is necessary to have rapid and reliable diagnostic methods that determine the risk factors associated with the diseases being studied. However, many of the tasks necessary in the health area such as clinical risk prediction, diagnosis, and therapy are more challenging for artificial intelligence applications (Maddox et al., 2019) (Bica et al. 2020).

Data from patients can be obtained from medical records, which generally are complex collections of data, with many stored variables, interactions between them, and often with a strict validity time. Additionally, the description of risk factors in the patient´s database is usually not perfect or complete (Malik et al., 2007) and data sets are generally unbalanced. The determination of risk factors is a major challenge due to the large amount of data currently generated from genetic studies and clinical data obtained in the medical office of some hospitals.

Additionally, to offer a tool aimed to support experts in the medical field in making decisions, it is necessary to present the results obtained in a way that provides transparency about the underlying mechanisms of the systems that have been used to analyze data to generate knowledge (Doran et al., 2018).

To address these challenges, this work presents a methodology to determine the risk factors associated with complex diseases using the machine learning approach. The methodology was tested in two application scenarios: Age-Related Macular Degeneration (AMD) and Preeclampsia. It is important to note that the methodology includes everything from generating data to delivering an interpretable decision-making system.

Contenido

Índice de tablas

Índice de figuras

Introducción

En años recientes ha habido un importante progreso de la inteligencia artificial que ha llevado a su aplicación con éxito en muy diversos dominios como el financiero, vehículos autónomos, robótica, mercadotecnia y cuidado de la salud, entre otros. La habilidad de la inteligencia artificial y en particular el aprendizaje automático para identificar interacciones y patrones en grandes conjuntos de datos ha permitido que se utilicen con éxito para construir sistemas de toma de decisiones automáticas. Así, la capacidad de aprendizaje y proveer soporte a la toma de decisiones de la inteligencia artificial está revolucionando el futuro del cuidado de salud. La inteligencia artificial se ha usado para diagnóstico y pronóstico de enfermedades, optimización del tratamiento y predicción de resultados, desarrollo de fármacos y para hacer frente a los problemas de salud pública (Noorbakhsh-Sabet, N., Zand, R., Zhang, Y., & Abedi, 2019).

Sin embargo, incorporar en la práctica modelos de aprendizaje automático en los sistemas de toma de decisión clínicos implica lidiar con retos que son más desafiantes que en otros dominios (Bica, I., Alaa, A. M., Lambert, C., & van der Schaar, 2020; Maddox, T. M., Rumsfeld, J. S., & Payne, 2019). Estos retos están relacionados tanto desde el punto de vista de los datos como de los modelos, algunos de estos se explican a continuación.

El área médica tiene particularidades que deben tomarse en cuenta, tales como la dificultad para obtener datos de los pacientes atendidos por los médicos. Para ello es necesario seguir protocolos de ética aprobados por las autoridades correspondientes y los hospitales (Centro del Conocimiento Bioético, 2015). Adicionalmente, para los estudios destinados al campo médico suele ser necesario contar con información de personas que padecen cierta enfermedad y con la de individuos que no la padecen. En donde, no necesariamente se cuenta con igual número de personas para cada caso. La naturaleza de los datos médicos y de salud es muy variada. Son frecuentemente no estructurados, multimodales y pueden contener potenciales sesgos (Reddy, S., Fox, J., & Purohit, 2019; Xiao, C., Choi, E., & Sun, 2018).

La medicina personalizada hace énfasis en la prevención, el diagnóstico temprano y el tratamiento de las enfermedades para evitar complicaciones y lograr los mejores resultados para los pacientes. Esto implica que existe la necesidad de métodos de tamizaje, diagnóstico

y pronóstico rápidos y confiables para el reconocimiento temprano de los pacientes en riesgo de padecer enfermedades para llevar a cabo la terapia adecuada para ellos o para ordenar y realizar más exámenes médicos necesarios (Bach et al., 2017).

Para resolver esta necesidad se requiere de detectar los factores de riesgo asociados a las enfermedades bajo estudio, esta información se puede obtener de los registros médicos. Los registros médicos son colecciones específicas de datos con gran complejidad, con una gran cantidad de variables almacenadas, con interacciones entre los atributos, que a menudo suceden en un marco de tiempo estricto para la validez de entradas médicas particulares. Así mismo, es evidente por la práctica médica que los datos de los pacientes según todos los factores de riesgo nunca son perfectos, ni completos (Malik et al., 2007).

Las bases de datos obtenidas a partir de los registros médicos a menudo son desequilibradas, multidimensionales e incluso redundantes. Esto suscita el problema del desbalanceo en las clases. Por esta razón, la preparación de los datos a analizar generalmente es necesaria para mejorar la calidad de los datos almacenados en las bases de datos médicas y para permitir una mejor extracción del conocimiento significativo (Bach et al., 2017).

Así, los algoritmos que funcionan con un desempeño satisfactorio para conjuntos de datos balanceados no lo hacen para los que tienen número desigual de datos para una y otra clase (H. Guo & Viktor, 2004). La exactitud alcanzada por un algoritmo puede ser adecuada para dominios distintos al de la medicina. Sin embargo, en el ámbito médico puede representar peligro de dejar sin atención médica a un individuo enfermo. Por lo tanto, es necesario implementar técnicas de aprendizaje automático que mejoren el desempeño de los clasificadores cuando trabajan con conjuntos de datos no balanceados (Baldi et al., 2000).

Desde el punto de vista de los modelos encontramos el problema de transparencia, explicabilidad e interpretabilidad. En el dominio médico, la presentación de los resultados obtenidos debe ser comprensible. Estos deben exponerse de forma que los expertos médicos puedan dar seguimiento a la forma en que se ha llegado a los resultados con el fin de que confíen en el sistema de soporte a toma de decisiones. La presentación de resultados interpretables deberá hacerse de manera que se dé transparencia en los mecanismos subyacentes de un sistema que se ha ocupado de analizar los datos para generar conocimiento (Doran et al., 2018).

De esta manera, en el dominio médico se requiere de una metodología para el tratamiento y análisis de los datos, desarrollado con base en las características y retos particulares anteriormente mencionados tales como la claridad en los mecanismos seguidos para obtener los resultados. Este método debe incluir la construcción de los conjuntos de datos, en donde se incluyan los datos pertinentes para su análisis. Además de abordar los problemas del tratamiento adecuado de estos para mejorar el entendimiento de las enfermedades bajo estudio.

Si bien, existen otros trabajos en los que se han abordado algunos de los pasos propuestos en esta metodología. Sólo en el caso de van der Schaar (Alaa & van der Schaar, 2018) se han aplicado varios de los pasos propuestos en esta metodología, para predecir la supervivencia a corto plazo de los pacientes con fibrosis quística. En este trabajo se pre procesan los datos, se aplican modelos para la clasificación, se menciona el problema del impacto del desbalanceo de las clases, y se presentan los resultados por medio de reglas extraídas de los modelos para explicar la predicción de supervivencia que separa a los pacientes que están realmente en riesgo de aquellos que no necesariamente necesitan un trasplante de pulmón a corto plazo. Así, en el trabajo de van der Schaar no se incluye la construcción de la base de datos, no se trata el problema del desbalanceo en las clases y se presentan los resultados de manera interpretable sólo a nivel general. En comparación con la metodología propuesta en este trabajo, el estudio de van der Schaar tiene un abordaje parcial del problema.

En consecuencia, la presente investigación propone una metodología para la determinación de factores de riesgo para enfermedades complejas, que integre la generación de datos, el diseño de las bases de datos, el procesamiento que permita solventar los problemas generados a partir de la escasez de los datos adecuados para lograr conjuntos de datos con cantidades balanceadas de datos de sujetos que sufren la enfermedad y sujetos sanos, la clasificación de los conjuntos de datos y la presentación interpretable de los resultados obtenidos.

Una metodología con estas características puede servir como una herramienta de apoyo para que el personal médico diagnostique y pronostique algunas enfermedades complejas de forma que estas puedan tratarse a tiempo o prevenirse.

Justificación

Este trabajo se justifica tanto desde el punto de vista de aprendizaje automático como desde el punto de vista médico.

Desde el punto del aprendizaje automático, sabemos que hay trabajos reportados en la literatura que abordan o se enfocan en alguno de los retos específicos en las diferentes fases del proceso de la creación de un sistema de soporte de toma de decisiones, sin embargo, es necesaria una metodología que ayude a tomar en cuenta retos con respecto a los datos y modelos específicos de este dominio (Alaa & van der Schaar, 2018).

Desde el punto de vista médico, las muertes causadas por enfermedades complejas podrían evitarse si se éstas se previenen y diagnostican de manera temprana. El reto que representan las enfermedades multifactoriales es desarrollar una metodología de identificación de individuos de riesgo en la población. Una mejor prevención y diagnóstico tendrán contribuciones sociales y económicos que se detallan a continuación.

En particular, para el caso de Degeneración Macular Relacionada con la Edad se busca proveer una herramienta para la prevención y diagnóstico temprano en mexicanos, que es una población insuficientemente estudiada respecto a esta enfermedad.

Para el caso de Preeclampsia, este trabajo busca apoyar a los médicos en la prevención y diagnóstico de la enfermedad con base en datos obtenidos en una población diferente a la mexicana, pero con alta prevalencia a nivel mundial.

Lo anterior se ha hecho con el fin de que los médicos, como usuarios finales de este trabajo, puedan recurrir a él para tomar de decisiones informadas, así como para contar con elementos para solucionar controversias en el diagnóstico y pronóstico de enfermedades complejas.

a) Justificación social

La prevalencia de las enfermedades complejas en la población a nivel mundial aumenta cada día debido, entre otras causas, a la longevidad de la población. Las enfermedades complejas representan una severa carga para el sistema de salud de todos los países. Esto es originado por el alto costo del acceso a los avances médicos con altas

especificaciones técnicas, y a la intervención muy tardía, cuando ya no puede obtenerse un beneficio significativo para la salud del paciente (David et al., 2006).

La identificación de variantes genéticas en las enfermedades complejas puede aportar conocimiento para avanzar en nuevos descubrimientos biológicos que permitan el desarrollo de biomarcadores mejorando la terapéutica y prevención de grupos de riesgo.

En el estudio de las enfermedades complejas, el aprendizaje automático ofrece apoyo en el pronóstico y diagnóstico. El aprendizaje automático aprende las reglas y patrones a partir de los datos, de manera inversa a como lo hace un médico en la práctica. Comenzando con las observaciones a nivel del paciente, los algoritmos filtran un gran número de variables, buscando combinaciones que puedan predecir los resultados de manera confiable. Los algoritmos de aprendizaje automático son capaces de manejar de enormes cantidades de predictores que combinan de manera no lineal y altamente interactiva (Mullainathan & Spiess, 2017). Esta capacidad nos permite utilizar nuevos tipos de datos, cuyo volumen o complejidad hace que en otro tiempo resultara inimaginable analizarlos.

b) Justificación económica

El diagnóstico clínico de las enfermedades complejas requiere de pruebas y estudios que resultan de alto costo para los pacientes, su familia y los sistemas de salud. Adicionalmente, la carencia de salud genera otros costos como los costos de atención informal, los costos de productividad, los costos de viaje de los pacientes y los gastos de bolsillo como el cuidado de niños, asistencia domiciliaria y compra de medicamentos (Leal et al., 2006).

Si bien, los algoritmos predictivos no pueden eliminar la incertidumbre médica, mejoran la asignación de recursos de atención de los pacientes. Los sistemas de alerta temprana que antes habrían tardado años en crearse ahora pueden desarrollarse y optimizarse rápidamente a partir de datos reales. La combinación del software de aprendizaje automático con el mejor equipo médico permitirá la entrega de atención que supera lo que cualquiera de los dos puede hacer solo. Los recursos de información y datos se deben usar con el fin de mejorar de forma sistemática nuestra salud colectiva (Chen & Asch, 2017).

Cada vez más, la capacidad de transformar los datos en conocimiento interviene en el campo de la medicina. El aprendizaje automático mejora el pronóstico de enfermedades. Los modelos de pronóstico (por ejemplo, la puntuación de Fisiología aguda y Evaluación de salud crónica [APACHE (Wagner & Draper, 1984), por sus siglas en inglés Acute *Physiology And Chronic Healt Evaluation*] y la Puntuación de evaluación de falla orgánica secuencial [SOFA (Lopes Ferreira et al., 2001), por sus siglas en inglés *Sequential Organ Failure Assessment*] se restringen a un número limitado de variables debido a que los humanos deben ingresar y contar las puntuaciones. Mejores pronósticos de enfermedad podrían mejorar la planificación de la atención para los pacientes con enfermedades complejas. El aprendizaje automático también plantea como posible apoyo para mejorar el diagnóstico de enfermedades (Obermeyer & Emanuel, 2016).

Es innegable que el camino de la prevención y la promoción de la salud es la opción adecuada por razones humanitarias y sociales, además de por razones económicas. Una posible disminución de los problemas económicos provocados por las enfermedades complejas puede ser la prevención de ellas por medio de la prognosis apoyada por los métodos desarrollados en aprendizaje automático.

c) Justificación tecnológica

Actualmente es común usar una metodología para desarrollar cualquier sistema inteligente de soporte a toma de decisiones que implique las siguientes fases: preparación de datos, selección de variables, creación de modelos y evaluación de estos. Sin embargo, las aplicaciones en el campo del cuidado de la salud conllevan retos particulares en la creación e integración de bases de datos, preparación de los datos y, en particular, en la importancia de que los modelos puedan ser explicables e interpretables a nivel del modelo y del individuo (Bica et al., 2020).

Los datos necesarios para determinar factores de riesgo para las enfermedades complejas se obtienen de las personas que acuden a la consulta médica. Obtener los datos adecuados, esto es, los de personas que padecen la enfermedad y sujetos sanos suele ser un proceso largo. Al ser datos provenientes de personas es necesario cumplir con estrictos protocolos de manejo ético. Adicionalmente, los registros de datos para cada individuo cuentan con inconsistencias que es necesario depurar.

De manera consecuente, los datos médicos normalmente presentan gran desbalanceo de clases, lo que se presenta comúnmente en los estudios de casos y controles. En donde, los datos de los sujetos que no padecen la enfermedad bajo estudio son de difícil acceso.

Adicionalmente, los expertos del ámbito médico se muestran desconfiados ante los resultados obtenidos, pues estos no se presentan de forma en la que se explique con claridad cómo se ha llegado a las inferencias logradas. Los métodos actuales están hechos para tener un tipo de interpretación por ejemplo para definir los factores de riesgo a nivel global o para definir los factores para un paciente o para entender la interacción entre los factores. Es necesario tomar todos estos factores en cuenta.

Por todo lo anterior resulta necesaria una metodología integral especialmente diseñada para resolver problemas del dominio médico que permita proporcionar una herramienta de apoyo para la toma de decisiones para los médicos, en donde su adopción se procure a través de la presentación de resultados de manera entendible vigilando los tres niveles de interpretabilidad.

Objetivo general

Desarrollar una metodología para apoyar al personal médico en el diagnóstico y pronóstico de enfermedades complejas para encontrar las posibles asociaciones entre estas y sus factores de riesgo resolviendo los retos propios de los datos médicos y de interpretabilidad del modelo.

Objetivos particulares

1. Desarrollar una metodología para apoyar al personal médico en el diagnóstico y pronóstico de enfermedades complejas para encontrar las posibles asociaciones entre enfermedades complejas y sus factores de riesgo resolviendo los retos propios de los datos médicos y de interpretabilidad del modelo.
2. Construir bases de datos que incluyan factores de riesgo y variantes genéticas a partir de muestras de sangre obtenidas en la consulta médica.
3. Aplicar técnicas de remuestreo para las bases de datos generadas con el fin de afrontar el problema del desbalanceo en las bases de datos generadas.
4. Aplicar técnicas de eliminación de datos espurios generados en el remuestreo.

5. Determinar los factores de riesgo más relevantes para la clasificación del conjunto de datos de casos y controles para algunas enfermedades complejas.
6. Presentar los resultados obtenidos de las predicciones aplicando técnicas de interpretabilidad para mejorar el entendimiento de los resultados generales y de instancias particulares.

Tabla A. Diseño de la investigación

Preguntas	Hipótesis	Objetivos
P1. ¿Cómo se pueden integrar las fases del proceso de creación de un sistema de soporte para la toma de decisiones en el dominio del cuidado de la salud?	H1. Con una metodología integral de aprendizaje automático enfocada en los retos particulares del dominio de cuidados de la salud que proporcione una herramienta de apoyo confiable para el diagnóstico y pronóstico de enfermedades complejas.	Desarrollar una metodología para apoyar al personal médico en el diagnóstico y pronóstico de enfermedades complejas para encontrar las posibles asociaciones entre enfermedades complejas y sus factores de riesgo resolviendo los retos propios de los datos médicos y de interpretabilidad del modelo.
P2. ¿Cómo se puede resolver el problema de la escasez de los datos adecuados para la clasificación?	H2. Se puede afrontar el problema de la escasez de datos adecuados para la clasificación por medio del remuestreo con la aplicación de técnicas de remuestreo.	Construir bases de datos que incluyan factores de riesgo y variantes genéticas a partir de muestras de sangre obtenidas en la consulta médica. Aplicar técnicas de remuestreo para las bases de datos generadas con el fin de afrontar el problema del desbalanceo en las bases de datos generadas. Aplicar técnicas de eliminación de datos espurios generados en el remuestreo.
P3. ¿Cómo se pueden determinar las variables más relevantes que indiquen los factores de riesgo para algunas enfermedades complejas?	H3. Las técnicas de selección de variables permiten determinar los factores de riesgo asociados con algunas enfermedades complejas.	Determinar los factores de riesgo más relevantes para la clasificación del conjunto de datos de casos y controles para algunas enfermedades complejas.
P4. ¿Las técnicas de interpretabilidad permitirán presentar los resultados de las predicciones de manera clara y accesible a los expertos en el área médica?	H4. Los resultados de las predicciones se pueden presentar en forma entendible para los médicos aplicando las técnicas de interpretabilidad.	Presentar los resultados obtenidos de las predicciones aplicando técnicas de interpretabilidad para mejorar el entendimiento de los resultados generales y de instancias particulares.

Fuente: Elaboración propia

Aportación al conocimiento de la investigación propuesta

Este trabajo propone una metodología para apoyar al personal médico en el diagnóstico y pronóstico de algunas enfermedades complejas, con el fin de encontrar las posibles asociaciones entre ellas y sus factores de riesgo para afrontar los desafíos propios de los datos médicos y de interpretabilidad del modelo.

El propósito de este trabajo es integrar en una metodología que incluya: la creación de los conjuntos de datos, la detección de los factores de riesgo, el balanceo de los conjuntos de datos y la presentación de resultados de manera entendible para el personal médico.

Los trabajos previos han abordado los pasos enumerados de manera separada. Una metodología especialmente diseñada para datos del dominio médico permitirá apoyar a los usuarios finales con un procedimiento completo e interpretable para la toma de decisiones para llegar al pronóstico y diagnóstico de enfermedades complejas.

La mayoría de los trabajos previos que estudian el problema de las clases no balanceadas han dado prioridad a la evaluación del desempeño de los clasificadores. En las tareas enfocadas al ámbito médico las bases de datos con frecuencia contienen pocos datos útiles y estos no están balanceados, lo que provoca que se logre bajo desempeño en la clasificación correcta de la clase minoritaria. El costo de la clasificación errónea es muy alto en términos de salud y riesgo. Así mismo, es necesario explicar con claridad al personal médico los patrones obtenidos por el algoritmo de aprendizaje automático para hacer comprensibles los resultados.

De esta manera, en este trabajo se presenta una metodología basada en las técnicas de aprendizaje automático para encontrar posibles asociaciones entre algunas enfermedades complejas y sus factores de riesgo con los siguientes elementos: construcción de la base de datos útiles para el pronóstico y/o diagnóstico de enfermedades complejas, balanceo de clases por medio de técnicas de sobremuestreo y eliminación de datos espurios; prueba y selección de algoritmos de clasificación adecuados para los datos; presentación de resultados en forma interpretable, considerando el nivel de datos, de algoritmos y de predicciones.

Organización de la tesis

Este trabajo está compuesto por cuatro capítulos. En el primer capítulo se presenta el marco teórico conceptual, en donde se hace una revisión general del aprendizaje automático, como parte de la Inteligencia Artificial, del problema de los conjuntos de datos con clases no balanceadas, sus probables causas, y las estrategias que se han usado tratar de solucionar el problema. Así mismo, en ese capítulo se presenta una descripción sobre el problema de aplicación de algoritmos de aprendizaje automático en el apoyo al diagnóstico y pronóstico en el área de la medicina. En el segundo capítulo se presenta la metodología empleada para desarrollar el modelo metodológico propuesto para abordar el problema desde el enfoque del aprendizaje automático. En el capítulo tres se presentan los resultados obtenidos al aplicar el modelo metodológico propuesto, que se enfoca en dos conjuntos de datos del dominio médico. Finalmente, se presentan las conclusiones derivadas de este trabajo.

Capítulo 1 Métodos de análisis de datos y aprendizaje automático en el diagnóstico médico

En el presente capítulo se detallan los principales conceptos del aprendizaje automático para la clasificación de conjuntos de datos con clases no balanceadas, la aplicación de estos conceptos en enfermedades complejas, y la interpretabilidad de los resultados en el dominio médico.

1.1 Conceptualización de la Inteligencia Artificial y el aprendizaje automático

El término inteligencia artificial (IA) tiene varias definiciones posibles según el contexto y las aplicaciones que se le den. Como disciplina académica, se considera que fue fundada en 1956 en la conferencia de Dartmouth (Moor, 2006). La IA es una rama de las ciencias computacionales encargada de estudiar modelos de cómputo capaces de realizar actividades propias de los seres humanos con base en dos de sus variables primordiales: el razonamiento y la conducta (López Takeyas, 2012).

El aprendizaje automático es una disciplina científica del ámbito de la Inteligencia Artificial que crea sistemas que aprenden automáticamente. En este contexto, aprender significa identificar patrones complejos a partir de datos (González, 2014).

Los algoritmos de aprendizaje automatizado se agrupan según la forma en que aprenden y sus resultados, estos son principalmente aprendizaje supervisado y no supervisado.

El aprendizaje supervisado es una técnica para deducir una función a partir de datos de entrenamiento. El fin de este tipo de aprendizaje es construir un modelo con base en los datos previamente clasificados. El clasificador resultante se usa para clasificar datos con clases desconocidas (Kotsiantis, 2007).

El aprendizaje no supervisado se ocupa de los datos que no se han clasificado previamente y no tienen un atributo de clase asociado con ellos. Es pertinente mencionar que el aprendizaje no supervisado es un enfoque de aprendizaje donde las instancias se colocan automáticamente en grupos significativos según su similitud (Kotsiantis & Pintelas, 2004).

1.2 Procesamiento y análisis de los datos

En el aprendizaje supervisado de datos obtenidos de algunos dominios como la medicina, se presenta el problema del desequilibrio de clase. En tal problema, casi todas las instancias están etiquetadas como una clase, mientras que muchas menos de ellas están etiquetadas como otra clase, generalmente la clase más importante. En este caso, los algoritmos estándar de aprendizaje automático tienden a verse influenciados por la clase mayoritaria e ignoran la clase minoritaria, ya que los clasificadores tradicionales buscan un rendimiento preciso en una amplia gama de casos (S. Kotsiantis et al., 2010).

Un ejemplo de lo antes mencionado se presenta en el contexto médico. El problema del desequilibrio de clases en los conjuntos de datos tiene importancia desde el punto de vista del aprendizaje automático, tanto como en el de la medicina.

El paso previo para el análisis de los datos es su preprocesamiento, en donde los conjuntos de datos se examinan para identificar los valores faltantes para sustituirlos por la moda o la media si son categóricos o numéricos, respectivamente.

Una vez pre procesados los conjuntos de datos, se procede a seleccionar las variables más relevantes.

1.2.1 Selección de las variables más relevantes

Con el fin de seleccionar las variables que resultan más importantes para la clasificación, se hace una selección de ellas sobre todas las variables del conjunto de datos. Para lograrlo, se utiliza el algoritmo de eliminación recursiva de variables (*RFE*, por sus siglas en inglés *Recursive Feature Elimination),* que ajusta el modelo a todas las variables.

Se usa el algoritmo *Random Forest* para medir la calidad de cada combinación de variables. Cada variable se clasifica según la importancia para el modelo. La precisión de los modelos generados es la medida de evaluación para determinar el poder predictivo del conjunto de variables en el proceso de clasificación.

El algoritmo RFE ajusta el modelo a todas las variables presentes en el conjunto de datos. Después cada variable se clasifica según la importancia para el modelo. Posteriormente, el algoritmo crea modelos utilizando S_i variables, con $i = 1 \ldots S$.

El algoritmo RFE intenta todas las combinaciones posibles y se mantiene en una lista de combinación de variables y su rendimiento. Para cada iteración, todas las variables se clasifican nuevamente. Al final de la ejecución del algoritmo, se elige la cantidad de variables que logre el mejor rendimiento de este.

Una vez que se han seleccionado las variables más importantes para los conjuntos de datos, se hace frente al problema del desbalanceo de las clases a través de sobre muestreo y submuestreo.

En el ámbito de aprendizaje automático, los algoritmos de clasificación tienen dificultades al clasificar bases de datos desbalanceadas, produciendo resultados erróneos. Se genera el problema del sobre entrenamiento, lo que provoca sesgos en los resultados hacia la clase mayoritaria (Ren et al., 2017). Esto tiene implicaciones trascendentes, pues las predicciones obtenidas con base en ellos no son fiables, puesto que se generan resultados en los que se incrementa el número de falsos positivos y falsos negativos especialmente. En el campo de la medicina, los datos obtenidos en la consulta hospitalaria son, generalmente costosos y difíciles de obtener. Al tratarse de datos provenientes de personas, puede resultar complejo obtener el consentimiento de los pacientes.

Así mismo, la recolección de datos resulta complicada porque depende de factores como el tiempo dedicado al registro de información, o bien, el manejo de muestras de tejidos para obtener datos a partir de ellas. Por esta razón, buscar como solución aumentar el tamaño de los conjuntos de datos no es una opción viable. Es necesarios explorar otras posibilidades que permitan trabajar con los datos existentes, para solucionar el problema. A continuación, se presentará el problema desde ambos puntos de interés.

1.3 Revisión del problema causado por las clases no balanceadas

El problema del desequilibrio de clases ha recibido atención significativa en áreas como el aprendizaje automático y el reconocimiento de patrones en los últimos años. Un conjunto de datos de dos clases está implícitamente desequilibrado cuando una de las clases está muy poco representada en contraste con la otra (García et al., 2012). Esta preocupación es esencial en aplicaciones del mundo real donde es costoso clasificar erróneamente ejemplos de la clase minoritaria, como la detección de llamadas telefónicas fraudulentas, el diagnóstico de

enfermedades, la recuperación de información, la categorización de texto y tareas de filtrado (Sotoca et al., 2007).

Los algoritmos canónicos de aprendizaje automático suponen que el número de objetos en las clases consideradas es más o menos similar. Sin embargo, en muchas situaciones de la vida real, la distribución de ejemplos es sesgada ya que los representantes de algunas clases aparecen con mucha más frecuencia. Esto plantea una dificultad para el aprendizaje de algoritmos, ya que estarán sesgados hacia el grupo mayoritario (Fernández et al., 2013). Al mismo tiempo, generalmente la clase minoritaria es la más importante desde la perspectiva de la minería de datos porque, a pesar de su escasez, puede generar conocimientos significativos y útiles (Krawczyk, 2016).

1.3.1 **Causas del problema del desequilibrio en las clases de los conjuntos de datos**

El desempeño de los clasificadores estándar tiende a disminuir al aplicarlos a conjuntos de datos no balanceados (Zhu et al., 2019). Las posibles causas de esta disminución son la existencia de clases poco representadas, la falta de densidad en los datos de entrenamiento, el traslape entre clases, y la confusión originada por los datos espurios (Krawczyk, 2016). Estas causas se enlistan a continuación:

- La existencia de clases poco representadas (*Small Disjuncts*): las clases poco representadas en los conjuntos de datos pueden ser confundidas con datos atípicos o espurios. Lo que ocasiona que la convergencia de la clase menos representada sea más lenta, disminuyendo la capacidad de generalización del clasificador.
- Falta de densidad en los datos de entrenamiento (*Lack of density*): al no encontrarse una zona en el espacio de atributos con suficiente densidad, se merma la capacidad de los métodos para inducir un patrón.
- Traslape entre clases (*Class Separability Problem*): puede aparecer un solape entre las clases de las instancias en las zonas limítrofe entre clases. Dando lugar a que ambas clases tengan una representación similar en estos tramos (Denil & Trappenberg, 2010), lo que provoca que sea imposible separar ambas.

- Confusión con ruido (*Noisy data*): los registros atípicos, como pueden ser los datos que están dañados o distorsionados, tienen especial relevancia en los conjuntos de datos no balanceados. Esto es debido a dificultad para discriminarlos frente a los sobrerrepresentados (Soft computing and intelligent information systems (SCI2S), 2020).

1.3.2 Soluciones al problema del desbalanceo en las clases

Las soluciones al problema de los conjuntos de datos con clases desbalanceadas se abordan desde el nivel de datos y desde el nivel de los algoritmos.

1.3.2.1 Solución a nivel de datos

Un problema de clasificación binaria desequilibrada tiene un conjunto de $x_i \in \mathbb{R}^d, y_i \in \{-,+\}, i = 1 \ldots., (N_+ + N_-)$ entrenamiento con N_+ como la clase minoritaria y N_- como la clase mayoritaria de los datos $N_+ \ll N_-$. Las técnicas de clasificación convencionales sobre ajustan naturalmente la clase mayoritaria ya que las muestras mayoritarias están excesivamente representadas en la función de pérdida. Como las técnicas de regularización convencionales están preparadas para equilibrar el sesgo y la varianza, no logran regularizar los sesgos unilaterales, como el sobreajuste de una clase en detrimento de la otra (Kovács, 2019).

Los métodos de nivel de datos para equilibrar las clases consisten en volver a muestrear el conjunto de datos original, ya sea sobre muestreando la clase minoritaria o sub muestreando la clase mayoritaria, hasta que las clases estén aproximadamente igualmente representadas. Ambas estrategias han mostrado inconvenientes importantes. El submuestreo puede arrojar datos potencialmente útiles, mientras que el sobre muestreo aumenta artificialmente el tamaño del conjunto de datos y, en consecuencia, empeora la carga computacional del algoritmo de aprendizaje. Sin embargo, ambos métodos han sido criticados principalmente por alterar la distribución original de la clase (Ramyachitra & Manikandan, 2014).

El método más simple para aumentar el tamaño de la clase minoritaria corresponde al sobre muestreo aleatorio, es decir, un método no heurístico que equilibra la distribución

de la clase a través de la replicación aleatoria de ejemplos positivos. Sin embargo, este método puede aumentar la probabilidad de sobreajuste, ya que hace copias exactas de las instancias de clase minoritaria.

Otra estrategia consiste en aplicar una técnica para sobre muestrear la clase minoritaria y, en lugar de simplemente replicar casos pertenecientes a la clase minoritaria, generar nuevas instancias sintéticas minoritarias al interpolar entre varios ejemplos positivos que se encuentran muy juntos (Blagus & Lusa, 2013). Aunque el sobre muestreo aumenta el costo computacional del algoritmo de aprendizaje, los experimentos realizados por Batista et al. (Batista et al., 2004), muestran la conveniencia de aplicar esta técnica cuando el conjunto de datos tiene muy pocos ejemplos positivos (minoritarios), en donde los métodos de sobremuestreo proporcionan resultados más precisos que los métodos de submuestreo considerando el área bajo la curva AUC.

El submuestreo aleatorio tiene como objetivo equilibrar el conjunto de datos mediante la eliminación aleatoria de ejemplos negativos. A pesar de su simplicidad, se ha demostrado empíricamente que es uno de los métodos de muestreo más efectivos.

El principal problema de esta técnica es que puede descartar datos potencialmente importantes para el proceso de clasificación (Bach et al., 2017). A diferencia del método aleatorio, muchas propuestas se basan en una selección más inteligente de los ejemplos de clase mayoritaria a eliminar (Vluymans, 2019). Algunas investigaciones indican la conveniencia de aplicar las estrategias de sub muestreo cuando el nivel de desequilibrio es considerado muy bajo (Sotoca et al., 2007).

Otro método para balancear las clases es la técnica de sobre muestreo de minorías sintéticas (SMOTE, por sus siglas en inglés: *Synthetic Minority Oversampling Technique*), que genera ejemplos sintéticos, operando en el espacio de atributos en lugar de en el espacio de datos. La clase minoritaria se sobre muestrea tomando cada muestra de clase minoritaria e introduciendo ejemplos sintéticos a lo largo de los segmentos de línea que unen a cualquiera de los k vecinos más cercanos de la clase minoritaria. Este método originalmente fue propuesto para conjuntos de datos con valores continuos (Chawla et al., 2002), con el paso del tiempo se han presentado variantes a la propuesta original que son capaces de manejar conjuntos de datos nominales y continuos, conocidos como SMOTE-N (Chawla, 2002).

Mohammed (2020) ha aplicado el método para balancear datos provenientes de registros médicos de pacientes que padecen diabetes, en donde los resultados muestran que el método de remuestreo y las técnicas de normalización tuvieron un efecto positivo en el rendimiento del modelo de clasificación, usando como métricas *Accuracy, Recall, F1, Presicion y ROC*.

Los ejemplos sintéticos hacen que el clasificador cree regiones de decisión más grandes y menos específicas. La clase mayoritaria se sub muestrea eliminando al azar muestras de la población de la clase mayoritaria hasta que la clase minoritaria se convierta en un porcentaje específico de la clase mayoritaria. Los resultados del uso de SMOTE en una simulación muestran que el enfoque SMOTE puede mejorar la precisión de los clasificadores para una clase minoritaria. La combinación de SMOTE y sub muestreo funciona mejor que el sub muestreo simple (Fernández et al., 2018).

1.3.3 **Eliminación de las instancias con ruido**

Los datos categóricos o nominales son usados para nombrar o categorizar información. Este tipo de datos no están ordenados, incluso si se usan números para representarlos. Por esta razón es necesario transformar los datos de manera que cada categoría de la variable tratada tenga un valor numérico, que no interfiera con la clasificación. Finalmente, múltiples columnas contienen la información de una variable, previamente codificada de manera binaria, esto es, con 0 y 1. Esto se realizó por medio del algoritmo de codificación *"One hot"*.

Una vez generadas las variables correspondientes, es deseable eliminar las instancias de clases opuestas que son sus propios vecinos más cercanos, es decir, que están muy cercanas entre sí. Para hacer esta tarea, se aplicó el algoritmo de Tomek (Batista et al., 2004), que busca esos pares de instancias y elimina la instancia mayoritaria de cada par.

El objetivo de este algoritmo es aclarar la frontera entre las clases minoritarias y mayoritarias, haciendo que las regiones minoritarias sean notablemente distintas a las de la clase mayoritaria para favorecer el trabajo de clasificación.

1.3.3.1 Soluciones a nivel algorítmico

De acuerdo con Zhu et al., (2019) existen diversas propuestas para abordar el problema del desequilibrio desde un punto de vista algorítmico, esto es, adaptar los algoritmos y técnicas existentes a las variables especiales de los datos desequilibrados. En este grupo están los algoritmos de aprendizaje con enfoque sensible al costo, los clasificadores de una clase y los ensambles de clasificadores o clasificadores múltiples (Z. H. Zhou, 2012).

1.3.3.2 Enfoque sensible al costo

Los modelos de aprendizaje tradicionales suponen implícitamente los mismos costos de clasificación errónea para todas las clases (Khan et al., 2018). Sin embargo, en algunos dominios, el costo de un tipo particular de error puede ser diferente de otros. El objetivo del enfoque sensible al costo es reducir el costo de los ejemplos clasificados incorrectamente en vez de los errores de clasificación. Algunos trabajos asignan costos distintos a los errores de clasificación para ejemplos positivos y negativos (Sun et al., 2015). En otros casos se propone el uso de costos de error no uniformes definidos por medio de la relación de desequilibrio de clase presente en el conjunto de datos (G. M. Weiss & Provost, 2003). Otros autores emplean un método para normalizar los costos de error en términos del número de ejemplos en cada clase (Liu & Zhou, 2006).

En aplicaciones reales, los costos de clasificación errónea generalmente son desconocidos, y en muchos casos su estimación es muy complicada, ya que dependen de factores propios de su dominio (Chawla, 2010). A pesar de esto, existen estrategias para fijar los valores de los costos de los falsos negativos (CFN) y los costos de los falsos positivos (CFP). En el caso de CFP, generalmente se asigna el valor 1, lo que se considera un costo mínimo (McCarthy et al., 2005). Para la aplicación en el dominio médico, donde un FP corresponde a un sujeto diagnosticado como enfermo cuando no lo está, se puede asignar un costo mínimo a la predicción incorrecta, ya que por medio de pruebas diagnósticas más complejas podrían encontrar el error. En el caso del CFN existe una estrategia que asigna el costo de acuerdo al desequilibrio entre clases (Japkowicz & Stephen, 2002). Esto podría ser insuficiente, ya que un FN es un paciente diagnosticado como sano cuando realmente está enfermo, con consecuencias de alto riesgo.

En general, los costos de clasificación errónea pueden describirse mediante una matriz de costos arbitraria *C*, siendo *C (i, j)* el costo de predecir que un ejemplo pertenece a la clase *i* cuando en realidad pertenece a la clase *j*. Los elementos diagonales generalmente se establecen en cero, lo que significa que la clasificación correcta no tiene costo (ver Tabla 1.1).

Tabla 1.1 Matriz de costos para un problema de clasificación binaria.

CVP	CFP
CFN	CVN

Fuente: Elaboración propia.

Las soluciones a nivel algorítmico y los enfoques sensibles al costo dependen más de los problemas, mientras que el nivel de datos y los enfoques de ensambles de métodos de clasificación basados en el procesamiento de datos son más versátiles (Díez-Pastor et al., 2015).

Las métricas mencionadas funcionan adecuadamente para conjuntos de datos balanceados.

1.3.4 **Ensambles de clasificadores**

Otra opción para abordar el problema de las clases no balanceadas es la de los ensambles, que se ha definido como un conjunto de clasificadores entrenados individualmente cuyas decisiones se combinan al clasificar nuevos objetos. El ensamble de clasificadores es una línea de investigación bien establecida, principalmente porque se ha observado que la precisión predictiva de una combinación de clasificadores independientes supera a la del mejor clasificador único (Tamez, 2018; Corso & Gibellini, 2011).

La construcción del modelo de clasificación se hace a través de un modelo tipo ensamble, que incluye a varios algoritmos de clasificación con el fin de mejorar el desempeño de estos. Existen varias formas para construir los ensambles: i) Modelos de ensambles secuenciales*, ii)* Modelos de ensambles en paralelo.

i) Los modelos de ensambles secuenciales aprovechan la dependencia entre los algoritmos de aprendizaje. Un algoritmo usualmente aceptado para hacer ensambles secuenciales es *Boosting,* que entrena secuencialmente a un grupo

de algoritmos de clasificación combinándolos para hacer predicciones. El término B*oosting* se refiere a un grupo de algoritmos que transforman algoritmos de clasificación débiles en robustos. Así, los clasificadores finales se centran más en los casos que fueron mal clasificados por los clasificadores anteriores en la secuencia (Galar et al., 2012). El algoritmo general de *Boosting* se muestra en la Figura 1.1.

Figura 1.1 Algoritmo de Boosting

Entrada: Distribución de los datos ($\mathcal{D}$);

Algoritmo base de aprendizaje $\mathcal{L}$;

Número de rondas de aprendizaje T

Proceso:

1. $\mathcal{D} = \mathcal{D}_i$ *% Se inicializa la distribución*
2. *Para t=1, ..., T:*

   $h_i = \mathcal{L}(\mathcal{D}_t)$*; %Se entrena un algoritmo de clasificación débil para la distribución* D_t

   $\epsilon_t = P_{x\sim\mathcal{D}_t}\big(h_t(x) \neq f(x)\big)$; *% Evaluación del error de* h_t

   $\mathcal{D}_{t+1} = Ajuste\ de\ la\ distribución\ (\mathcal{D}_t, \epsilon_t)$

   Fin
3. *Salida:* $H(x) = sign\ (\sum_{t=1}^{T} h_t(x))$

ii) Los modelos de ensamble paralelo aprovechan la independencia entre los algoritmos de clasificación. El error puede reducirse combinando algoritmos base independientes entre sí. Al tomar una clasificación binaria como referencia, la función elemental es *f,* y cada clasificador base tiene un error de generalización ϵ_t, para cada clasificador base h_i, se tiene que

$$P\big(h_i(x) \neq f(x)\big) = \epsilon \qquad (1)$$

Al combinar T veces cada clasificador base, el ensamble H tendrá error sólo cuando al menos la mitad de los clasificadores tengan errores (ver ecuación 2).

$$H(x) = signo\,(\sum_{i=1}^{T} h_i(x)\,) \qquad (2)$$

La generalización del error en el ensamble, de acuerdo la desigualdad de Hoeffding clásica referida por Avni et al. (2019) se muestra en la ecuación 3, en donde se identifica que la generalización del error se reduce exponencialmente para el ensamble de tamaño T. El error se aproximará a cero en la medida en que T aumente.

$$P\big(h_i(x) \neq f(x)\big) = \sum_{k=0}^{\lceil T/2 \rceil} \binom{T}{k}\,(1-\epsilon)^k\,\epsilon^{T-k} \leq \exp\left(-\tfrac{1}{2}\,T\,(2\,\epsilon-1)^2\right) \qquad (3)$$

El algoritmo de *Bagging* (Bootstrap Aggregating) (Galar et al., 2012) hace uso de la distribución de *Bootstrap* para subconjuntos de datos para entrenar a los algoritmos base. Dado un conjunto con *m* ejemplos de entrenamiento, se generará un conjunto de datos de tamaño *m* por muestreo con reemplazo.

Algunos de los ejemplos originales aparecerán más de una vez, mientras que algunos ejemplos originales no estarán presentes en la muestra. Repitiendo el proceso T veces, se obtendrán T conjuntos de datos con *m* ejemplos de entrenamiento. *Bagging* adopta la estrategia de "votos" para clasificación y "promedios" para la regresión.

En clasificación, para predecir una instancia en un conjunto de datos de prueba, el algoritmo alimenta a los algoritmos de clasificación y recolecta sus resultados, entonces se vota por las etiquetas y toma la etiqueta ganadora como la predicción. En la Figura 1.2 se muestra el algoritmo bajo el que trabaja *Bagging.*

Figura 1.2 Algoritmo de Bagging

Entrada: Distribución de los datos $\mathcal{D}$;

Algoritmo base de aprendizaje $\mathcal{L}$;

Número de rondas de aprendizaje T

Proceso:

1. $\mathcal{D} = \mathcal{D}_i$ *% Se inicializa la distribución*
2. *Para t=1, ..., T:*

   $h_t = \mathcal{L}(\mathcal{D}, \mathcal{D}_{bs})$; *% Se entrena un algoritmo de Bagging*

   $\epsilon_t = P_{x\sim\mathcal{D}_t}\big(h_t(x) \neq f(x)\big)$; *% Evaluación del error de h_t*
3. Fin

Salida: $H(x) = \underset{y\epsilon Y}{\operatorname{argmax}} \sum_{t=1}^{T} I(h_t(x) = y)$

1.3.5 **Ensamble apilado *(Stack)***

Un conjunto de los promedios de los modelos en un ensamble combina las predicciones de múltiples modelos entrenados. Una variación de este enfoque, llamada conjunto promedio ponderado, mide la contribución de cada miembro del conjunto en un conjunto de datos de reserva.

Esto permite que los modelos con buen rendimiento contribuyan con más y los modelos con menor rendimiento contribuyan con menos. El promedio ponderado del ensamble proporciona una mejora sobre el promedio conjunto del modelo.

Una generalización adicional de este enfoque es reemplazar la suma ponderada lineal del modelo (por ejemplo, regresión lineal) utilizado para combinar las predicciones de los submodelos con cualquier algoritmo de aprendizaje. Este enfoque se llama generalización apilamiento.

En el apilamiento, un algoritmo toma las salidas de submodelos como entrada e intenta aprender cómo combinar mejor las predicciones de entrada para hacer una mejor predicción de salida.

En este trabajo se prueban cinco modelos en un ensamble apilado:

i) Análisis discriminante lineal (LDA),
ii) Particionamiento recursivo y árboles de regresión (RPART),
iii) Regresión logística (a través del modelo lineal generalizado o GLM),
iv) Vecinos k-más cercanos (KNN),
v) *Support Vector Machine* con una función de núcleo de base radial (RSVM).

En un ensamble por apilamiento es deseable que las predicciones hechas por los submodelos tengan baja correlación (Brown, 2010). Esto significa que los modelos son diestros en distintos sentidos, lo que permite que un nuevo clasificador obtenga lo mejor de cada modelo para obtener un desempeño mejorado.

Si las predicciones para los submodelos estuvieran altamente correlacionadas (> 0.75), harían predicciones iguales o muy similares, reduciendo el beneficio de combinar las predicciones.

La diversidad es uno de los cimientos del buen funcionamiento de los ensambles. Un ensamble ideal debe tener clasificadores individuales precisos y sus errores deben estar en diferentes instancias. Los ensambles de métodos para el aprendizaje desequilibrado abordan el problema del desequilibrio de clases por medio de técnicas como la reponderación, el sobre muestreo y el submuestreo. Estas técnicas de preprocesamiento buscan entrenar a los clasificadores con un conjunto de datos menos desequilibrado. Estas técnicas de preprocesamiento no sólo abordan el problema del desequilibrio, sino que también agregan diversidad, ya que cada clasificador base está entrenado en una versión diferente del conjunto de datos. Los métodos de ensamble difieren según la forma en que inducen la diversidad entre los clasificadores base (Díez-Pastor et al., 2015).

El enfoque más común es modificar el conjunto de entrenamiento para cada miembro del ensamble. De esta forma, los ensambles que no han sido diseñados especialmente para los conjuntos de datos con clases desbalanceadas abordan el problema del desbalanceo combinando los modelos de clasificación con alguna de las técnicas de preprocesamiento. Esto se puede hacer combinando el sobre muestreo y submuestreo con algunos clasificadores. De otra forma, se aplican métodos diseñados para manejar conjuntos de datos no balanceados

(Díez-Pastor et al., 2015). Algunos de los métodos más usados para trabajar con conjuntos con clases desequilibradas son SMOTE (Chawla et al., 2002), SMOTE Boost (Chawla et al., 2003), ADASYN (Díez-Pastor et al., 2015) y BDSMOTE (Rivera & Xanthopoulos, 2016).

Algunas de las investigaciones realizadas para el problema de las clases desbalanceadas, proponen como trabajos futuros mejorar la tarea de clasificación de datos del ámbito médico, mejorando el desempeño de los clasificadores en términos de métricas de evaluación, sin una pérdida significativa en la interpretabilidad y fiabilidad de los modelos generados (Mena et al., 2012).

1.3.6 **Métricas de desempeño**

En general, el objetivo de un algoritmo de clasificación en aprendizaje automático es construir clasificadores que maximicen su exactitud. Sin embargo, esto no es suficiente para producir clasificadores robustos en problemas con conjuntos de datos no balanceados. La exactitud puede conducir a conclusiones erradas. Esto se debe a que ésta métrica considera la exactitud general y no la exactitud que corresponde a la clasificación de cada clase (Camaré, 2008).

La aplicación de algoritmos de aprendizaje automático sólo es exitosa y útil en campos como el diagnóstico médico si el dominio se entiende bien y se aplica un marco de evaluación adecuado (Jain et al., 2018). De esta manera, es necesario determinar cuál es la forma más apropiada de evaluar algoritmos de aprendizaje automático en problemas con conjuntos de datos no balanceados (X. Guo et al., 2008).

Para medir el desempeño de los algoritmos de clasificación usualmente se usa una matriz de confusión (G. Weiss et al., 2007), ver Tabla 1.2. A partir de ella se pueden calcular algunas métricas de evaluación tales como: *Accuracy*[1], sensibilidad, especificidad, precisión, valor pronóstico negativo y Kappa (ver ecuaciones 1- 8).

[1] Se usa el término inglés *Accuracy,* con el fin de evitar la ambigüedad entre exactitud y precisión en español.

Tabla 1.2 Matriz de confusión binaria

	Predicción Negativa	**Predicción Positiva**
Real Negativa	**VN**	**FP**
Real Positiva	**FN**	**VP**

Fuente: Elaboración propia.

Al aplicar algoritmos de aprendizaje automático sobre conjuntos de datos no balanceados, estos casi siempre producen clasificadores de alta *Accuracy y especificidad*, pero con baja *sensibilidad*. El objetivo principal de aprender de conjuntos de datos desequilibrados es mejorar la sensibilidad sin dañar la precisión. Por otro lado, los objetivos de sensibilidad y precisión a menudo pueden ser conflictivos, ya que al aumentar el número de verdaderos positivos para la clase minoritaria, también se puede aumentar el número de falsos positivos y esto reducirá la precisión (Chawla, 2010).

$$Accuracy\ total = \frac{VP + VN}{VP + VN + FP + FN} \tag{1}$$

$$Sensibilidad\ (Recall) = \frac{VP}{VP + FN} \tag{2}$$

$$Especificidad = \frac{VN}{VN\ +\ FP} \tag{3}$$

$$Precisión = \frac{VP}{VP + FP} \tag{4}$$

$$Valor\ pronóstico\ negativo = \frac{VN}{VN + FN} \tag{5}$$

$$Kappa = \frac{Accuracy\ total - Accuracy\ aleatoria}{1 - Accuracy\ aleatoria} \tag{6}$$

$$Accuracy\ aleatoria = \frac{(VN + FP) * (VN + FN) + (FN + VP) * (FP + VP)}{Total * Total} \tag{7}$$

$$Accuracy\ aleatoria = \frac{(Falsos\ reales) * (Predicción\ de\ falsos) + (Verdaderos\ reales) * (Predcción\ de\ verdaderos}{Total * Total} \tag{8}$$

El análisis ROC es un método estándar para evaluar el rendimiento de los clasificadores binarios (Saito & Rehmsmeier, 2015). El gráfico ROC muestra el equilibrio entre especificidad y sensibilidad. Aplica para todo el modelo porque muestra pares de valores de especificidad y sensibilidad calculados en todos los valores posibles del umbral. En las gráficas ROC, los clasificadores con un rendimiento aleatorio muestran una línea diagonal recta desde (0, 0) a (1, 1), esta línea se puede definir como la línea base de la curva ROC. Una curva ROC proporciona una medida de rendimiento única llamada Área bajo la curva ROC (AUC). Los puntajes de AUC son convenientes para comparar el desempeño de múltiples clasificadores (Berrar, 2019).

La media geométrica es una de las métricas de rendimiento estándar utilizadas en un clasificador de conjuntos de datos desequilibrados (Luque et al., 2019). Es una métrica con sesgo nulo si su enfoque está centrado en los éxitos de clasificación, no presenta ninguna limitación para la aplicación específica donde se utiliza. La razón para usarla es equilibrar la relación de predicción entre la clase mayoritaria y minoritaria. La proporción de esta métrica muestra que tan buen clasificador es el que predice las clases. Una ventaja de ROC y la media geométrica es que ambas combinan sensibilidad y especificidad *(*ver ecuaciones 2-3). En donde VP, FN, FP y VN se pueden explicar de la siguiente manera: Verdadero Positivo (VP) se refiere a la predicción correcta de la clase mayoritaria. Falso negativo (FN) se refiere a la predicción errónea de la clase minoritaria como la clase mayoritaria. Falso positivo (FP) se refiere a la predicción incorrecta de la clase mayoritaria como una clase minoritaria. Verdadero negativo (VN) se refiere a la predicción correcta de la clase minoritaria.

Las métricas de desempeño presentadas no son completamente eficientes para los conjuntos de datos desbalanceados. Con el fin de enfrentar este problema algunos autores han propuesto seleccionar las que mejor funcionen para el dominio en que se esté trabajando.

De esta manera, es necesario establecer cuáles son las métricas que mejor se ajustan mejor al dominio médico, que es al que la metodología propuesta en este trabajo será aplicada.

Accuracy es el porcentaje de clasificaciones correctas de todas las instancias. Es una descripción de errores sistemáticos, esto es, una medida del sesgo estadístico es una descripción de errores aleatorios. Si el valor de *Accuracy* es bajo, indica que hay una diferencia entre un resultado y su valor real.

Kappa es una métrica que compara la *Accuracy* esperada contra la observada. Se usa para evaluar a un clasificador o a varios clasificadores entre sí. Además, considera los aciertos debidos al azar. Es una medida de qué tan cerca están las instancias clasificadas de los datos etiquetados como verdaderos.

Kappa es un mejor indicador de cómo se desempeñó el clasificador en todas las instancias, que *Accuracy*. Esto se debe a que *Accuracy* puede sesgarse si la distribución de clases está similarmente sesgada (Landis & Koch, 1977).

Adicionalmente, el valor de Kappa de un modelo es directamente comparable con el de cualquier otro modelo utilizado para la misma tarea de clasificación.

Los ensambles de clasificadores han sido propuestos como una buena alternativa para el manejo de conjuntos de datos desbalanceados. Sin embargo, la métrica que se usa frecuentemente es la precisión. Esto puede mejorarse aplicando a los ensambles de clasificadores las técnicas de comparación de métricas propuestas para clasificadores individuales.

1.4 Presentación de resultados: interpretabilidad

En años recientes, los sistemas de IA y aprendizaje automático han logrado un alto rendimiento en muchas tareas que anteriormente se consideraba computacionalmente inalcanzables (LeCun et al., 2015). Se han hecho estos progresos en el campo de la IA debido al aumento de la información disponible y las mejoras de hardware y a la optimización de los algoritmos. Un reto para la adopción de la IA es entender cómo se llegó a una propuesta en particular, lo que es decisivo para la aceptación de los usuarios como elementos que respalden la toma de decisiones.

Con una mayor difusión de las aplicaciones de IA, es factible que los problemas relacionados con la confianza se conviertan en problemas más apremiantes. La confianza se refuerza con criterios específicos, pero existen problemas importantes con la incompletitud

en la formalización del problema. Dado que los criterios de confianza son difíciles de formalizar y cuantificar, los criterios de interpretabilidad y explicabilidad generalmente se utilizan como objetivos intermedios. En una etapa posterior, las explicaciones del sistema se pueden verificar para determinar si satisfacen los criterios de confianza deseables (Dosilovic et al., 2018).

La interpretabilidad en el contexto del aprendizaje automático es la capacidad de explicar o presentar en términos comprensibles para los humanos. La interpretabilidad y la explicabilidad a menudo se usan indistintamente. Algunos investigadores hacen distinción entre ellas.

De acuerdo con Montavon (2018), la interpretación es el mapeo de conceptos abstractos en un dominio que los humanos pueden tener sentido, mientras que la explicación es la colección de variables de un dominio, que han contribuido para que un ejemplo dado produzca una decisión.

Edwards et al., (2017) dividen las explicaciones en nociones centradas en el modelo y en el sujeto, que corresponden a las definiciones de interpretabilidad y explicabilidad de Montavon. De la misma forma, Doshi-Velez (2017) plantea la diferencia entre la interpretabilidad global y la local.

Aunque a menudo es imposible que una explicación sea completamente fiel a menos que sea la descripción completa del modelo, para que una explicación sea significativa debe ser al menos localmente fiel. Esto es, debe corresponder a cómo se comporta el modelo en la vecindad de la instancia predicha.

La fidelidad local no implica fidelidad global: las variables que son globalmente importantes pueden no ser importantes en el contexto local, y viceversa. Un explicador debería ser capaz de exponer la forma en la que trabaja cualquier modelo y, por lo tanto, ser independiente de él. Además de explicar las predicciones, es importante proporcionar una visión global para establecer la confianza en el modelo.

Con esta perspectiva, el modelo LIME (*Local Interpretable Model-agnostic Explanations*) propone un enfoque modular y extensible para explicar fielmente las predicciones de cualquier modelo de manera interpretable. Sobre la base de las explicaciones

para las predicciones individuales, selecciona algunas explicaciones para presentar al usuario, de modo que sean representativas del modelo (Ribeiro et al., 2016).

Se busca que el explicador sea capaz de replicar el comportamiento del modelo de manera local e interpretable. Para lograr este LIME minimiza la medida de infidelidad (ver ecuación 9).

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \ \mathcal{L}(f, g, \pi_x) + \Omega(g) \tag{9}$$

En donde:

f: predictor original

x: variables originales

g: modelo de explicación que podría ser un modelo lineal, árbol de decisión o listas de reglas

π: medida de proximidad entre una instancia de z a x para definir la localidad alrededor de x.

ξ: medida de infidelidad o pérdida local de g al aproximar f en la localidad definida por π.

Ω: medida de la complejidad del modelo de explicación g

Con el fin de garantizar tanto la interpretabilidad como la fidelidad local, se minimiza la pérdida de infidelidad al modelo, mientras se mantiene el segundo término de la ecuación suficientemente bajo como para que los usuarios puedan comprenderlo. De esta manera, mientras se optimiza para la pérdida local, LIME logra la fidelidad local.

Si g es el modelo para aprender, z es una instancia de los datos de entrenamiento y $f(z) = y$ entonces, para crear un conjunto de entrenamiento completo se realiza un muestreo aleatorio uniforme de x y se crean múltiples z desde una sola instancia de x. Los conjuntos de entrenamiento luego son ponderados por $\pi(x)$ para enfocarse más en las z que están más cerca de x.

Dado este conjunto de datos y etiquetas, la medida de infidelidad está optimizada para aprender el modelo de explicación. Esto significa que no hay dependencia del tipo de modelo original para que LIME proporcione explicaciones (agnósticas al modelo).

LIME trabaja de acuerdo con el algoritmo que se presenta en la Figura 1.3, en donde K es un límite en el número de variables a considerar para la explicación. A continuación, se hace que Ω tienda al infinito si $size(w) > K$. LIME utiliza explicadores lineales para aproximar el límite de decisión del modelo original.

Figura 1.3 Algoritmo de LIME

Entrada: Clasificador f, Número de muestras N

Entrada: Instancia x, y su versión interpretable x´

Entrada: Núcleo de similitud π_x*, tamaño de la explicación K*

$z \leftarrow \{\}$

$para\ i \in \{1,2,3,\dots,N\}\ \ hacer$

$z_i^{'} \leftarrow muestra\_cercana\_a(x^{'})$

$z \leftarrow z \cup \langle z_i^{'}, f(z_i), \pi_x z_i \rangle$

fin

$\omega \leftarrow K - Lasso(z, K) \rhd con\ z_i^{'}\ como\ variables, f(z)\ como\ la\ clase$

$regresa\ \omega$

De esta manera, se concluye que LIME es una buena aproximación local, independiente de los modelos de clasificación, que permite dar claridad a la forma en que se llega a los resultados para las instancias seleccionadas.

En concordancia con lo anterior, es evidente que no existe una definición universalmente aceptada del concepto de interpretabilidad. Sin embargo, es indispensable procurar presentar el trabajo hecho por el aprendizaje automático de manera que resulte inteligible para los usuarios de sus resultados.

La noción de interpretabilidad depende del público objetivo, de acuerdo con sus variables, deberá ser la explicación que se les deba presentar (Varshney et al., 2018).

En la práctica, existen tres categorías para la interpretabilidad (S.-M. Zhou & Gan, 2008):

I. La de los datos,
II. La de los algoritmos aplicados para llegar a las predicciones y
III. La de las predicciones.

La primera se refiere a qué datos se usaron para entrenar el modelo, sus variables y la explicación de por qué se usaron.

La interpretabilidad a nivel de algoritmo sólo requiere el conocimiento del algoritmo y no de los datos o el modelo aprendido. Algoritmos como el método de mínimos cuadrados para modelos lineales están bien estudiados y entendidos, se caracterizan por una alta transparencia. Sin embargo, los algoritmos de aprendizaje profundo son menos transparentes.

La tercera se refiere a la explicación de cómo se llegó a las predicciones propuestas. Permite entender qué algoritmos fueron aplicados, así como los umbrales que se tomaron para llegar a las predicciones planteadas. Aquí se explica cómo el algoritmo aprende un modelo a partir de los datos y qué tipo de relaciones puede aprender.

En el aprendizaje automático, a menudo se debe hacer una compensación entre precisión e interpretabilidad. Los modelos más precisos, como *Boosted trees*, *Random Forest* y las redes neuronales generalmente no son inteligibles. Los modelos más inteligibles, como la regresión logística, los Naive-Bayes y los árboles de decisión única, frecuentemente tienen una precisión menor. Aunque estas resultan útiles para hacer comprensibles para los usuarios, las predicciones obtenidas.

Esta compensación a veces limita la precisión de los modelos que se pueden emplear en aplicaciones de misión crítica, como la atención médica, donde es importante poder comprender, validar, editar y confiar en un modelo aprendido (Sturm et al., 2015).

1.5 Problemática para la aplicación de aprendizaje automático en el diagnóstico y pronóstico médico

La extracción de conocimiento de los conjuntos de datos del dominio médico es una tarea compleja. Los conjuntos de datos médicos son conocidos por su composición, en términos de ruido, valores perdidos y distribución de clases desequilibrada (Jain et al., 2018). Un problema importante por resolver es la selección de las variables relevantes para identificar los elementos de la clase de interés. Estos son conocidos como síntomas o factores de riesgo, para el diagnóstico y pronóstico en medicina respectivamente. Adicionalmente, estos son clasificados como modificables y no modificables. Los primeros son los que pueden apoyar en el diagnóstico de enfermedades, ya que los no modificables, como la edad, podrían

no resultar útiles para la intervención médica ya que no existe tratamiento para alterarlos. Esto puede complicar adicionalmente la tarea de clasificación.

En la investigación médica la mayoría de los datos se recolectan a partir de estudios observacionales, prospectivos y longitudinales. Estos se basan en la observación de una enfermedad en un grupo de pacientes durante un período determinado para establecer asociación entre los posibles factores de riesgo y la enfermedad (National Institutes of Health, 2014).

Con base en lo anterior, se realiza una clasificación binaria de los sujetos sanos y enfermos, dependiendo de si desarrollaron o no la enfermedad. Sin embargo, estos estudios sean diseñados para hacerse en un tiempo determinado lo cual complica la tarea de clasificación (Sedgwick, 2014). Como consecuencia, se tiene que un sujeto que presentó factores de riesgo durante el período de estudio podría morir de una causa distinta a la enfermedad estudiada o al finalizar el estudio podría no presentar todavía la enfermedad. En los dos casos el sujeto es clasificado como sano, condición que induce ruido, lo que confunde a los clasificadores. Dada esta problemática, el aprendizaje automático aplica técnicas que buscan allanar los posibles efectos, como el ruido.

El problema del diagnóstico y pronóstico médico resuelto por medio de técnicas de aprendizaje automático debe ser capaz de proveer al personal médico de un nuevo punto de vista acerca de la enfermedad en estudio. El clasificador debe revelar de forma comprensible la estructura de los patrones encontrados para llegar a los resultados evitando la pérdida de precisión en el proceso de clasificación. Los resultados deben ser comprensibles a nivel general y también a nivel local, considerando que los datos analizados provienen de personas a las que es necesario dar seguimiento individual.

Un campo fértil para la aplicación del aprendizaje automatizado es el estudio de las enfermedades complejas. Por sus características, implican un reto para el desarrollo de los métodos de aprendizaje automático adecuados para su correcto pronóstico y diagnóstico. A continuación, se caracteriza a las enfermedades complejas y se explica su problemática de estudio.

Capítulo 2 Estudios previos acerca de Enfermedades Complejas bajo el enfoque de Aprendizaje automático

2.1 Enfermedades complejas

Buena parte de las enfermedades más prevalentes en la población son el resultado de la combinación de factores hereditarios y ambientales. Muchos trastornos frecuentes, como la diabetes, la degeneración macular relacionada con la edad, la preeclampsia o la hipertensión arterial se presentan grupos familiares, lo que refleja su componente hereditario (aunque pueden existir también factores ambientales compartidos). Los estudios de epidemiología genética muestran que, a diferencia de las enfermedades hereditarias clásicas, en estos trastornos el riesgo no se explica por la variación sólo de un gen. A partir de esto surge la denominación de enfermedades poligénicas o complejas (Riancho, 2012).

Las enfermedades complejas son la principal causa de muerte en el mundo. Aproximadamente el 80% de las muertes por enfermedades complejas se producen en países de bajos y medianos ingresos (Londoño, 2017). Sin embargo, las muertes causadas por estas enfermedades podrían disminuirse si se implementaran programas de prevención y diagnóstico temprano, con estrategias orientadas al monitoreo poblacional o a la generación de modelos predictivos causales para detección temprana, interpretando las causas primordiales que generan la patología (Brunotto & Zárate, 2012).

Considerar las interacciones gen-ambiente puede mejorar la comprensión de las causas de las enfermedades complejas, y puede ayudar a los investigadores a desarrollar terapias dirigidas. En lugar de estudiar los factores genéticos y ambientales por separado, los investigadores ahora están estudiando cómo los genes y los factores ambientales interactúan entre sí (Craig, 2008).

Para entender las enfermedades complejas, primero es necesario conocer los principios de herencia de Mendel, que explican cómo los rasgos heredados pasaron de generación en generación. Estos son el principio de segregación y el principio de distribución independiente.

Existen muchas circunstancias que violan las reglas de herencia de Mendel, lo que puede dificultar la determinación de los factores que influyen en las enfermedades genéticas

complejas. Estos fenómenos incluyen penetrancia reducida, expresividad variable, definición del fenotipo (en los casos en que una enfermedad puede ser causada por más de un gen), rasgos poligénicos, interacciones gen-gen e interacciones gen-ambiente (Eichler et al., 2010).

Las enfermedades complejas no obedecen al patrón de herencia mendeliano dominante o de gen único recesivo. A pesar de esto, los estudios de enfermedades de un solo gen pueden proporcionar información sobre la contribución de los genes individuales a los fenotipos asociados con diversas enfermedades o condiciones complejas (Yang et al., 2010).

Tradicionalmente, la probabilidad y las técnicas estadísticas se han utilizado para comprender enfermedades complejas. La dificultad de manejar grandes cantidades de datos debido a su volumen, velocidad y variabilidad hace que su estudio sea factible computacionalmente, con la posibilidad de encontrar predictores relacionados con el inicio de la enfermedad es una razón clave para usar técnicas de aprendizaje automático (DeWan et al., 2006).

Los estudios de los factores de riesgo asociados con enfermedades complejas se han basado en procedimientos de modelos estadísticos tradicionales, que no tienen en cuenta la posibilidad de interacciones entre los factores de riesgo y el riesgo de enfermedad, ni la posibilidad de múltiples etiologías (Sing et al., 1992). Es necesario diseñar estudios que consideren las posibles interacciones entre los factores de riesgo para enfermedades complejas.

El enfoque estadístico tradicional para modelar enfermedades comunes asume que los atributos son independientes. Además, este enfoque supone que el mismo modelo puede explicar a todas las personas que padecen una enfermedad común. Estas suposiciones no son adecuadas debido a la complejidad biológica de estas enfermedades. En un modelo poligénico de herencia de la enfermedad, existe una relación de muchos a uno entre los genes y la enfermedad.

El uso de las técnicas de aprendizaje automático apoya en el enfoque de medicina personalizada, en donde se apuesta a la prevención de las enfermedades al detectar patrones y relaciones entre los padecimientos y sus factores de riesgo.

En este trabajo se incluyen dos enfermedades complejas de interés para el dominio médico y para el del aprendizaje automático: Degeneración Macular Relacionada con la Edad (DMRE) y Preeclampsia.

En los dos casos, se trata de enfermedades que implican riesgos importantes para la población. La DMRE provoca ceguera irreversible si no se da el tratamiento adecuado en fases tempranas de la enfermedad; la Preeclampsia, puede llegar a ocasionar la muerte de las mujeres que la padecen si no se diagnostica y atiende a tiempo. Ambas se presentarán como casos de uso en el capítulo 3.

2.1.1 **La Degeneración Macular Relacionada con la Edad (DMRE)**

La Degeneración Macular Relacionada con la Edad (DMRE) es la causa principal de disfunción visual y ceguera en los países desarrollados y una causa creciente en los países subdesarrollados. En Estados Unidos, su prevalencia en la población mayor de 65 años es del 9% y aumenta al 28% en los mayores de 75 años (Wong et al., 2014). En México, la prevalencia de la pérdida de visión en 2016 fue de 1,241,000 personas. En donde el glaucoma y la DMRE tienen un papel fundamental entre las condiciones no reversibles de pérdida de la visión (Jimenez-Corona & Graue-Hernandez, 2018).

La DMRE se caracteriza por una degeneración progresiva de la mácula, que causa pérdida de la visión del campo central. Un rasgo característico de la DMRE es la formación de depósitos en la mácula, llamados drusas, que pueden progresar hacia atrofia geográfica o neo vascularización sub retinal, indicadores de la DMRE tardía (Sivakumaran et al., 2011).

Los modelos de riesgo de DMRE se pueden agrupar en dos categorías: predicción e inferencia. En la primera categoría, el interés es desarrollar modelos que proporcionen el mejor rendimiento para la evaluación de riesgos. Los modelos predictivos de DMRE contienen una combinación de factores de riesgo genéticos, no genéticos y clínicos que podrían usarse para predecir el nivel de riesgo de un individuo. Los modelos predictivos no son totalmente aceptados por una parte de la comunidad médica. La Academia Americana de Oftalmología (AAO) ha declarado que las pruebas genéticas para enfermedades multifactoriales no serán una práctica de rutina hasta que los ensayos clínicos puedan demostrar que los pacientes con genotipos específicos se benefician de tipos específicos de terapia o vigilancia (Stone et al., 2012). La principal dificultad del desarrollo de una prueba

predictiva se basa en la complejidad de la enfermedad; los diversos fenotipos clínicos y otros factores de confusión que pueden limitar su utilidad. Hasta ahora, los modelos estadísticos clásicos han sido el principal método para modelar la DMRE.

Los estudios basados en factores de riesgo, especialmente las variantes genéticas, se han desarrollado recientemente debido a la exploración del genoma. A continuación, se presentan algunos de los trabajos más recientes enfocados en el estudio de factores de riesgo incluyendo variantes genéticas.

2.1.2 Enfoque de Aprendizaje automático para el estudio de Degeneración Macular Relacionada con la Edad

El propósito de las técnicas de aprendizaje automático es reconocer automáticamente patrones complejos en un conjunto de datos dado, permitiendo por lo tanto inferencia o predicción en nuevos conjuntos de datos (Duda & Stok, 2001). Esas técnicas se utilizan con frecuencia en las ciencias de la visión clínica. Los estudios realizados con ML han abordado el problema de las enfermedades de la retina bajo dos enfoques: el estudio de las imágenes de la retina y la asociación con factores de riesgo. En general, los estudios de asociación con factores de riesgo genéticos y ambientales tienen como objetivo principal el pronóstico de las enfermedades.

Dada la importancia del componente hereditario en las enfermedades complejas, los científicos han tratado de identificar los genes y los polimorfismos involucrados en las enfermedades. En los estudios de asociación de todo el genoma (GWAS, por sus siglas en inglés), se elige un gen considerado como un factor de riesgo (Martínez et al., 2011). Algunos de sus polimorfismos se identifican y analizan para determinar la asociación entre sus alelos y un fenotipo o la asociación con la frecuencia de una enfermedad.

Cada año, los GWAS se publican con un número creciente de asociaciones de polimorfismos de un solo nucleótido (SNPs) con enfermedades o fenotípicas (Hindorff et al., 2014). Se debe realizar un análisis estadístico adicional para encontrar un polimorfismo o variante de un gen asociado con una enfermedad en una población específica.

Primero, se estudia la agregación familiar para determinar si la enfermedad está determinada genéticamente. En segundo lugar, es necesario localizar genes de interés para la

enfermedad. En las áreas identificadas puede haber miles de polimorfismos de interés (Iniesta et al., 2005). Tradicionalmente, la probabilidad y las técnicas estadísticas se han utilizado para comprender enfermedades complejas.

La dificultad de manejar grandes cantidades de datos debido a su volumen, velocidad y variabilidad hace que los métodos actuales no sean factibles computacionalmente, pero la posibilidad de encontrar predictores relacionados con el inicio de la enfermedad es una razón clave para usar técnicas de aprendizaje automático (Zhang & Baird, 2016).

Uno de los objetivos principales de la medicina personalizada es identificar de manera pre sintomática a las personas con alto riesgo de enfermedad utilizando el conocimiento del perfil genético de la persona y sus factores de riesgo ambientales (Sobrin & Seddon, 2014), por lo que un pronóstico realizado a través de aprendizaje automático puede representar una herramienta de apoyo para los profesionales médicos (Castaneda et al., 2015).

La DMRE se ha estudiado bajo varios enfoques. Se ha explorado vastamente el estudio de imágenes obtenidas por medio de estudios de fondo de ojo, un enfoque menos explorado es el que toma en cuenta las variantes genéticas, además de otros factores de riesgo. En Martínez-Velasco et al., (2017) se ha presentado una revisión de los estudios realizados para encontrar los factores de riesgo asociados a la DMRE con ambos enfoques.

Bajo el enfoque de la medicina personalizada, Larrañaga (Larrañaga et al., 2006) menciona aprendizaje automático como una herramienta importante para transformar el enorme volumen de datos complejos en conocimiento. Este artículo presenta algunas de las técnicas más útiles para el modelado y optimización de la bioinformática y destaca en la aplicación de métodos de aprendizaje automático en genética.

Spencer et al.(2011) proponen aumentar los conjuntos de datos mediante el uso de la reducción de la dimensionalidad multi factorial (MDR) y la evolución gramatical de las redes neuronales (GENN), además del enfoque de regresión logística (LR). Combinando los resultados de los modelos LR y GENN, el algoritmo logra una sensibilidad del 77.0%, especificidad 74.1%.

Jiang et al. (2009) proponen una adaptación de bosque aleatorio (RF) para las interacciones epistáticas. Las principales contribuciones del algoritmo epi-Forest (detección

de interacciones epistáticas mediante la selección secuencial de variables de avance y RF) de ventana deslizante son la incorporación de la RF en estudios de control de casos y el cribado automatizado de los polimorfismos candidatos para un análisis estadístico. Los enfoques de aprendizaje automático se presentan como métodos de complemento para facilitar la exploración de interacciones entre múltiples polimorfismos, debido a que la epistasis desempeña un papel importante en la patogénesis de la DMRE. El autor propuso la importancia de Gini, obtenida a partir de métodos de clasificación de AM, puede complementar el valor p de los estudios estadísticos, para medir para las asociaciones entre polimorfismos y DMRE. Para identificar las posibles combinaciones de las variantes genéticas que son protectoras para DMRE.

Gold et al. (2006) analizaron un conjunto de factores de riesgo para DMRE con un modelo estadístico y a continuación, con un modelo de AM. Esto se hizo con un modelo basado en Algoritmos Genéticos (AG). La ventaja de los AG sobre los modelos tradicionales es su habilidad para incorporar múltiples loci a través del genoma para hacer una predicción. Esto permite a los modelos identificar las interacciones complejas entre polimorfismos, correlacionadas con sus predicciones.

Chen et al. (2007) propusieron un método basado en conjuntos de árboles, para identificar las interacciones entre genes y las interacciones entre genes y ambiente. Este método se propone para resolver el problema de los datos faltantes y para la selección de atributos simultáneamente. Este enfoque evita el problema de la colinealidad para datos de todo el genoma, y no requiere de ninguna suposición a priori. Los algoritmos basados en bosques son una herramienta de aprendizaje automático popular porque son adaptables a los datos, se aplica a problemas grandes y pequeños y es capaz de considerar la correlación y las interacciones entre las entidades. Los falsos positivos son una preocupación importante en la identificación de genes de enfermedades. Los autores demostraron que la tasa de falsos positivos del método puede distinguir con éxito regiones del genoma asociadas con la enfermedad de regiones neutrales con una tasa falsa – positiva (FPR) inferior al 5%.

Çelebiler (2013) estudió la relación entre la presencia de polimorfismos genéticos múltiples, factores de riesgo y DMRE seca y húmeda. Se construyeron tres tipos de redes bayesianas (BN) para investigar la relación entre la presencia de múltiples polimorfismos

genéticos y DMRE. Las redes bayesianas mejoran el proceso de aprendizaje mediante la fijación de conocimientos previos y causa menos ruido y sobreajuste que otros métodos. La flexibilidad de sus modelos permite tomar decisiones precisas a partir de datos inciertos.

Fraccaro et al., (2015) compararon los modelos de "caja blanca" (incluyendo regresión logística y árboles de decisión), como los más interpretables, y métodos de "caja negra" como máquina vectorial de soporte (SVM), *Random Forest y Boost* adaptativo. Ambos métodos identificaron las drusas blandas y la edad como las variables más importantes para diagnosticar la DMRE. El autor hace hincapié en la interpretación y la limitación del número de muestras necesarias para obtener resultados fiables para realizar diagnósticos tempranos. Se propone una interfaz gráfica de usuario que muestre la vía diagnóstica o la importancia variable para proporcionar a las especialistas vías de decisión para hacer que los diagnósticos tempranos sean factibles y diferenciar mejor los subconjuntos ambiguos de pacientes.

Krishnaiah et al., (2015) presentaron el rendimiento predictivo del modelo redes neuronales artificiales en comparación con la capacidad predictiva del modelo regresión logística, se muestra que las redes neuronales artificiales funcionan mejor debido a la relación entre variables (ver tabla 2.1).

Tabla 2.1. Trabajos previos para el estudio de DMRE bajo el enfoque de clasificación con factores de riesgo.

	Datos	**Tarea realizada**	**Método**	**Resultados**	**Consideraciones**
Spencer et al. (2011)	Variantes genéticas y conjunto de datos de factores de riesgo.	Identificar a las personas con alto riesgo de padecer DMRE.	MDR, Algoritmos genéticos, Regresión lineal	Sensibilidad 77.0%, Especificidad 74.1%	Conjuntos de datos de tamaño pequeño.
Jiang R. et al.(2009)	Conjunto de datos de polimorfismos.	Contribución de cada SNP a la clasificación.	Epi-Forest.	8.5% tasa de error de clasificación	Conjuntos de datos de tamaño pequeño.
Gold et al. (2006)	Conjunto de datos de polimorfismos y factores de riesgo	Identificar el riesgo y la protección que confieren algunos haplotipos para DMRE	Algoritmos genéticos.	Sensibilidad 58.37% Especificidad 77.13%	No toma en cuenta el tamaño y balanceo del conjunto de datos.
Chen X. et al. (2007)	Conjunto de datos de variantes genéticas	Identificar haplotipos relacionados con enfermedades.	Algoritmos de conjuntos de árboles.	Tasa de falsos positivos < 5%	Conjuntos de datos con más de 10 millones de instancias.
Çelebiler, A. et al.(2013)	Conjunto de datos de signos clínicos y variantes genéticas.	Relación entre polimorfismos y DMRE.	Redes de Bayes	Precisión 83.3% (±4.7)	Conjuntos de datos balanceados.
Fraccaro et al. (2015)	Conjunto de datos clínicos.	Obtener un Sistema interpretable para el diagnóstico de DMRE.	Regresión lineal, Arboles de decisión. Caja negra: Máquinas de soporte vectorial, *Random Forest* y *Boost* adaptativo.	Rendimiento medio Caja Blanca: 92%, Caja Negra: 90%.	Conjuntos de datos de tamaño pequeño.
Krishnaiah S. et al. (2015)	Conjunto de factores de riesgo.	Desarrollar modelos de predicción para DMRE.	Regresión lineal, Redes neuronales artificiales.	Sensibilidad 79% Especificidad 69%.	Desbalanceo en las clases.

Fuente: Elaboración propia

En algunos de estos trabajos se hacen consideraciones acerca de los conjuntos de datos a estudiar. En la mayor parte de ellos se considera la problemática de los conjuntos de datos pequeños. Se puede observar que sólo en una minoría se toma en cuenta el desbalanceo de clases.

La segunda enfermedad compleja estudiada en este trabajo es la Preeclampsia. A continuación, se explican sus características generales.

2.1.3 **Preeclampsia (PE)**

Una de las patologías más importantes que acompañan al embarazo es el síndrome de preeclampsia / eclampsia. El síndrome es un trastorno multisistémico que puede incluir cambios cardiovasculares, anomalías hematológicas, insuficiencia hepática y renal y manifestaciones neurológicas o cerebrales (Duckitt & Harrington, 2005). La preeclampsia es un síndrome clínico que afecta al 3-5% de los embarazos y es una de las principales causas de mortalidad materna, especialmente en los países en desarrollo. Es un trastorno hipertensivo multisistémico (Roberts et al., 2012).

El diagnóstico de PE se puede mejorar con el uso de métodos de salud electrónica. El enfoque actual de los investigadores de atención médica es promover el uso de la tecnología de salud electrónica en los países en desarrollo para apoyar las decisiones médicas (Zayyad & Toycan, 2018). Los factores socioeconómicos bajos actúan como factores de riesgo múltiples para la preeclampsia. En México, el bajo estatus socioeconómico de las mujeres duplicó el riesgo de preeclampsia y eclampsia (Cerón-Mireles et al., 2001). Un estudio en Australia encontró que las mujeres que trabajan en comparación con las que no trabajan tienen un mayor riesgo de desarrollar preeclampsia y eclampsia (Najman et al., 1989). Esto puede estar relacionado con el estrés que sufren las mujeres durante el trabajo. Las técnicas de aprendizaje automático se han utilizado para apoyar a los expertos en salud, en la prevención de preeclampsia (Neocleous et al., 2009).

2.1.4 **Enfoque de Aprendizaje automático para el estudio de Preeclampsia**

Martínez-Velasco et. al, (2018) presentaron una revisión de estudios relevantes para determinar los factores de riesgo asociados a preeclampsia, bajo el enfoque de aprendizaje automático. El objetivo de utilizar técnicas de aprendizaje automático en estos casos de estudio es detectar patrones y relaciones entre los factores de riesgo y la enfermedad. Los trabajos más relevantes para el estudio de la PE se enlistan a continuación (ver Tabla 2.2). Se incluye el enfoque de determinación de factores de riesgo y el enfoque de la biología molecular, en todos ellos se aplicaron métodos de aprendizaje automático.

Tabla 2.2. Estudios realizados con Aprendizaje automático para Preeclampsia

Autor/ año	Datos	Tarea realizada	Método	Consideraciones
Kenny et al. (2005)	Matriz de datos obtenida de una gráfica 3D. Contiene tres metabolitos pico en cada muestra.	Clasificación	Programación Genética basada en árboles.	Consideran bases de datos balanceadas. Muestran reglas para mejorar la interpretabilidad.
Neocleous et al., (2009)	La base de datos incluye 15 parámetros.	Clasificación	Redes neuronales con una estructura neuronal multi capa.	Conjuntos de datos pequeños.
Espinilla, (2017)	Conjunto de datos que contiene factores de riesgo.	Clasificación.	Arboles de decisión. Transformación lingüística difusa.	
Mackenzie et al., 2016	Conjunto de datos de exosomas extraídos del plasma.	La contribución de diferentes tejidos a la expresión génica.	Deconvolución (Factorización de Matriz Negativa).	No considera conjuntos de datos desbalanceados.
Cox et al., 2011	Proteína de membrana plasmática	Clasificación.	Se utilizaron varios algoritmos de aprendizaje automático.	No considera conjuntos de datos desbalanceados.
Velikova et al., (2013)	Factores de riesgo, signos e historial clínico.	Clasificación.	Redes Bayesianas.	No considera conjuntos de datos desbalanceados.
Tejera et al., (2011)	Conjunto de datos del historial clínico.	Clasificación.	Redes Neuronales Artificiales.	No considera conjuntos de datos desbalanceados.
Villa et al., (2017)	Conjunto de datos de factores de riesgo.	Agrupamiento.	Agrupamiento Bayesiano.	No considera conjuntos de datos desbalanceados.
Moreira et al., (2016)	Factores de riesgo, mecanismos fisiológicos, síntomas.	Clasificación.	Redes Bayesianas.	No considera conjuntos de datos desbalanceados.
Fergus et al., (2018)	Base de datos de polimorfismos.	Clasificación.	Autocodificadores apilados de aprendizaje profundo.	No considera conjuntos de datos desbalanceados.
Mehta et al., (2016)	Revisión y análisis de métodos de minería de datos aplicados al dominio de atención materna.			No aborda el problema de las clases desbalanceadas.

Fuente: Elaboración propia

En particular, Kenny (2005) propuso una programación genética basada en árboles para diagnosticar PE analizando tres metabolitos en plasma sanguíneo. Los autores consideran metabolitos en plasma obtenidos de 87 casos y controles, presentan algunos datos demográficos y analizan la concentración de algunos metabolitos en ambos grupos. Ellos presentan resultados en un cromatograma y algunas reglas para obtener conclusiones.

Neocleous (2009) propuso redes neuronales para clasificar una base de datos que contiene 15 factores de riesgo con los mejores resultados obtenidos fueron con una estructura neuronal multi capa para estimar el riesgo de aparición de PE en una etapa temprana.

También se presentan gráficos sobre el rendimiento de las redes neuronales utilizadas. Los resultados se informan en tablas y conclusiones por medio de textos.

Así mismo, Espinilla et al., (2017) clasificaron un conjunto de datos de factores de riesgo utilizando árboles de decisión sin poda y transformación difusa lingüística utilizando un enfoque genético. Este trabajo presenta una metodología que permite un seguimiento lingüístico en tiempo real. Presentaron algunas reglas generadas por un árbol de decisión.

Velikova et al.,(2014) propusieron una clasificación de base de datos por medio del modelo de redes bayesianas que se construye manualmente utilizando conocimientos de expertos en el tema. Los datos de entrada son los factores de riesgo y las mediciones de los signos se proporcionan en la aplicación móvil. El conjunto de datos contiene historial médico y algunos datos externos. Los autores presentan interfaces de usuario que muestran el riesgo de sufrir la enfermedad basada en la red bayesiana temporal. No presentan ningún elemento que respalde la interpretabilidad de los expertos médicos.

Tejera et al., (2011) clasificaron un conjunto de datos de historia clínica que incluye índices de variabilidad de la frecuencia cardíaca materna y factores de riesgo para caracterizar la PE. Los resultados se presentan en términos de curvas ROC, sensibilidad una variación de especificidad e importancia normalizada de las variables independientes en las redes neuronales artificiales obtenidas. No se presentan elementos para mejorar la interpretabilidad de los resultados.

Villa et al., (2017) propusieron agrupamiento bayesiano para calcular la relación de riesgo de cada enfermedad. El análisis indica un aumento exponencial en el riesgo de preeclampsia a medida que aumenta el número de factores de riesgo. El análisis se basa en un estudio de casos y controles y los registros médicos de ellos. Este trabajo muestra los resultados del análisis de clúster en el mapa de calor que presenta los factores de riesgo en los diferentes clústeres. La forma de tomar las decisiones del sistema no es explícita. Moreira et al., (2016) propusieron la clasificación de factores de riesgo, mecanismos fisiológicos y conjunto de datos de síntomas para identificar el embarazo de alto riesgo. La principal contribución de este trabajo incluye la presentación de una red bayesiana construida para ayudar a los responsables de la toma de decisiones en momentos de incertidumbre en el cuidado de las mujeres embarazadas. Este trabajo se centró en la construcción de un sistema

inteligente diseñado para apoyar una decisión médica para la atención médica de la mujer embarazada.

Fergus et al., (2018) clasificaron un conjunto de datos de variantes genéticas basado en el Estudio de Asociación de Genoma completo (GWAS) utilizando autocodificadores apilados de aprendizaje profundo para permitir la detección de PE. Propusieron incluir en futuras obras que utilizaran reglas lógicas estructuradas para reducir la interpretabilidad de los modelos de redes neuronales.

Cox et al., (2011) plantearon Redes de Bayes como el mejor algoritmo para clasificar un conjunto de datos de proteínas de membrana plasmática. Presentan exhaustivamente los procedimientos y conjuntos de datos utilizados.

Mehta et al.,(2016) presentaron una revisión y análisis de los métodos de minería de datos aplicados a la atención materna. En términos de interpretabilidad, los autores concluyen que la representación gráfica de los árboles de decisión y los modelos de Bayes ingenuos son más fáciles de entender, a diferencia de las redes neuronales y las máquinas de soporte vectorial.

Con base en los trabajos consultados, se concluye que los algoritmos de aprendizaje automático además de tener un buen desempeño deben presentar los resultados obtenidos de manera interpretable para ser capaces de ofrecer apoyo en la toma de decisiones. Esto se deberá respaldar con métricas que permitan medir la fiabilidad de los resultados obtenidos. Es indispensable considerar que los modelos obtenidos deberán trabajar eficientemente con pocos datos y, en muchos casos, con clases desbalanceadas.

Lo anterior es necesario porque en el dominio médico los datos son escasos y resulta costoso obtenerlos. De forma que es necesario diseñar una metodología para clasificación para los datos del dominio médico, que contemple los conjuntos de datos con clases desbalanceadas, que logre disminuir el número de falsos negativos en las predicciones y que, adicionalmente, resulte comprensible para el personal médico.

En este capítulo se presentó una revisión de los problemas ocasionados por el desbalanceo de las clases en los conjuntos de datos, los métodos utilizados para balancearlas y diversas propuestas de interpretabilidad. Se presentó un panorama general de las

enfermedades complejas, en particular se exponen los casos de la Degeneración macular relacionada con la edad y la Preeclampsia, pues estas enfermedades serán los escenarios de aplicación de la metodología propuesta en este trabajo.

En el siguiente capítulo se presenta la metodología desarrollada en este trabajo para tratar los conjuntos de datos del dominio médico, con el fin de generar un instrumento de apoyo en la toma de decisiones en el diagnóstico y pronóstico de algunas enfermedades complejas.

Capítulo 3 Estrategia Metodológica

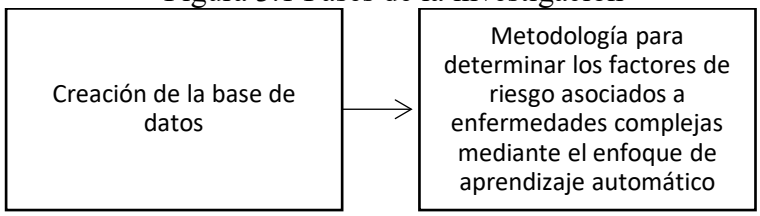
En este capítulo de describe de forma detallada el procedimiento para integrar la metodología propuesta que determinará los factores de riesgo asociados a enfermedades complejas

En la presente investigación se aborda el problema de la escasez de los datos útiles en el dominio médico. Se plantea el uso de técnicas de sobre muestreo y sub muestro para mejorar la tasa de desbalanceo en conjuntos de datos médicos. A continuación, se eliminan los datos espurios generados en el remuestreo, estos se clasifican por medio de ensambles. Los ensambles probados se evalúan por medio de la métrica seleccionada con base en su grado de consistencia. Finalmente, se presentan los resultados de la clasificación en forma interpretable a nivel global y local.

La metodología propuesta se realiza en dos fases consecutivas (ver Figura 3.1):

I. Creación de la base de datos con base en las muestras de sangre y datos obtenidos en la consulta médica hospitalaria.
II. Metodología para la determinación de los factores de riesgo y presentación de resultados interpretables dirigida a los usuarios finales.

Figura 3.1 Fases de la investigación



Fuente: Elaboración propia.

3.1 Creación de la base de datos

Los datos se obtuvieron a partir de muestras de sangre periférica de pacientes en la consulta médica hospitalaria. A partir de las muestras de sangre venosa se obtuvo el ADN (ácido desoxirribonucleico) para cada muestra. El ADN se genotipó en el laboratorio de

Biología Molecular con base en los polimorfismos de interés. Finalmente, se construyó la base de datos para estudiarla por medio de las técnicas de aprendizaje automático (ver Figura 3.2).

Los polimorfismos fueron seleccionados por medio de un GWAS (*Genomic Wide Association Study*), seguido de investigaciones previas en las que se han estudiado los polimorfismos asociados con la DMRE en poblaciones diferentes a la mexicana (Sivakumaran et al., 2011). Una vez seleccionados los polimorfismos de interés, se procede a la recolección de los datos.

3.1.1 **Procedimiento de recolección de datos**

La extracción de ADN se hizo por medio de la obtención de 3 ml de sangre venosa en tubos con anticoagulante (EDTA). La obtención de muestras se realizó con apego a los principios de la Declaración de Helsinki (World Medical Associationdeclaration of Helsinki, 2014) y se contó con consentimiento informado por escrito de todos los individuos incluidos en esta investigación.

El ADN se extrajo con el Kit PureGene DNA purification whole Blood (QIAGEN) bajo las siguientes condiciones: se colocaron 3 ml de Solución de Lisis de Eritrocitos (RBC) en un tubo de 15 ml. Se agregó 1ml de sangre y se mezcló por inversión. Se incubó durante 5 min a temperatura ambiente.

Terminado el tiempo de incubación se centrifugó a 13000 rpm durante 2 min. Se desechó el sobrenadante y agregó 1 ml de Solución de Lisis celular y 340 µl de solución de precipitación de proteínas, se mezclaron por agitación con vórtex y se centrifugaron las muestras a 13000 rpm.

Se recuperó el sobrenadante y se depositó en un tubo limpio de 15 ml, se agregó 1ml de isopropanol y se homogenizó por inversión para precipitar el ADN. Se centrifugó 3 min a 1300 rpm, se retiró el sobrenadante y se agregó 1 ml de etanol al 70% para lavar el ADN precipitado, se centrifugó 1 min a 1300 rpm.

El ADN se hidrató en 300 µl de agua inyectable y se almacenó a -20°C hasta su uso. Se midió la pureza y se cuantificó el ADN por espectrofotometría (nanodrop) y se valoró su integridad en geles de agarosa al 0.8% con GelRed (Biotim).

Una vez que se obtuvo el ADN de cada muestra, se procedió a la genotipación de del ADN de cada individuo.

3.1.2 **Ensayos de discriminación alélica**

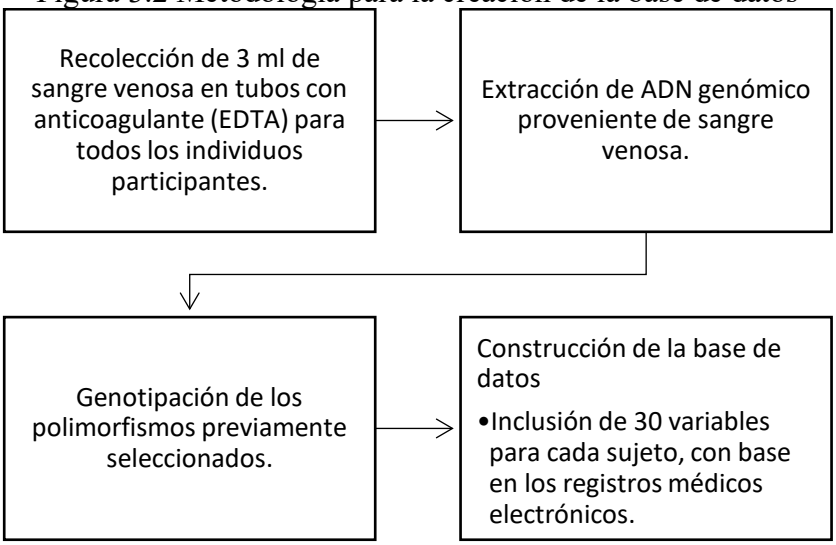
Para realizar la genotipación de los polimorfismos previamente seleccionados se usaron ensayos de discriminación alélica por sondas *Taqman* (Applied Byosistem). Para cada muestra se utilizaron 20 ng de ADN, 10 µl de máxima *Probe qPCR Master Mix 2X*, (Thermo Scientific), 24µl de *primers* y c.b.p 20µl volumen final de reacción. Todos los genotipos se determinaron en un instrumento de PCR Tiempo Real Piko Real (Thermo Scientific). Las sondas se adquirieron directamente de ensayos sobre demanda del servicio de Applied Byosistems como parte del control de calidad para la verificación del proceso de genotipificación. Todas las determinaciones se hicieron por duplicado.

3.1.3 **Construcción de la base de datos**

La base de datos se construyó mediante la recopilación de dos variantes genéticas (CFH1, CFH2) y variables demográficas para la DMRE obtenidas de los registros médicos electrónicos recopilados en la consulta oftalmológica de los participantes.

Las variables incluidas fueron: cataratas bilaterales, consumo de alcohol, pterigión, retinopatía diabética, glucosa alterada, diabetes, hemorragia vítrea, edad, obesidad, sexo, hipercolesterolemia, xerosis en ambos ojos, catarata del ojo derecho, presbicia, astigmatismo, tabaquismo, edema macular, blefaritis, desprendimiento vítreo posterior completo en el ojo derecho, ruptura coroidea, dislipidemia y ectropión.

Figura 3.2 Metodología para la creación de la base de datos



Fuente: Elaboración propia.

Ya que la base de datos ha sido construida, se sigue con la segunda fase de la metodología, esto es, la determinación de los factores de riesgo por medio de técnicas de aprendizaje automático.

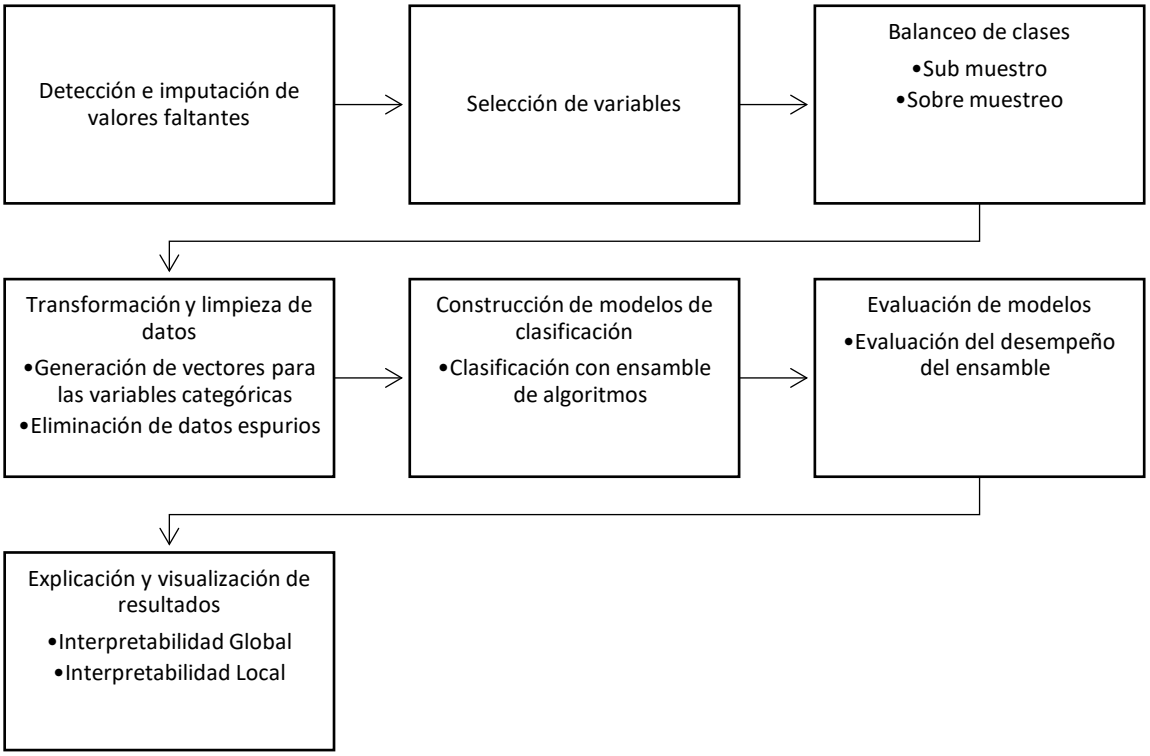
3.2 Determinación de los factores de riesgo por medio de técnicas de aprendizaje automático

Una vez que se cuenta con la base de datos, como primer paso, se procede a la detección de datos faltantes. Estos se manejan por imputación, las variables continuas se sustituyen por la mediana y las variables dicotómicas por la moda.

A continuación, se hace la selección de las variables más relevantes por medio de la eliminación recursiva de variables. Se mejora el equilibrio de las clases con la aplicación de técnicas de sobre muestreo y submuestreo de los conjuntos de datos.

Después, se soluciona el problema generado por las variables categóricas generando vectores de datos, inmediatamente se eliminan de las instancias espurias generadas por el proceso de sobre y submuestreo. Posteriormente, se hace la clasificación por medio de un ensamble de algoritmos, se evalúa su desempeño con las métricas adecuadas y se presentan los resultados en forma comprensible a nivel global y local (Figura 3.3).

Figura 3.3 Metodología para determinar los factores de riesgo asociados a enfermedades complejas mediante el enfoque de aprendizaje automático



Fuente: Elaboración propia.

3.2.1 **Selección de variables**

La selección de las variables más relevantes es el proceso de elegir un subconjunto de variables para su uso en la construcción de modelos. Las técnicas de selección tienen como objetivo la simplificación de modelos para que sean más fáciles de interpretar, disminuir los tiempos de entrenamiento y reducir la dimensionalidad de los datos. Las técnicas de reducción de dimensionalidad se han convertido en una necesidad obvia en el campo de la medicina cuando se hace uso de las técnicas de aprendizaje automático. Actualmente, se genera una gran cantidad de datos en el dominio médico. Esto incluye los síntomas que puede tener un paciente y también muchos informes de pruebas médicas que pueden generarse. (Khalid et al., 2014).

Los datos con alta dimensionalidad representan complicaciones para los algoritmos de clasificación debido al alto costo computacional y el uso de memoria (Janecek et al., 2008).

La ventaja de la selección de variables es que no se pierde información sobre la importancia de ninguna de las variables. Pero tiene la desventaja de que se requiere un conjunto reducido de variables y las originales son muy diversas, la información se puede perder ya que algunas de las variables deben omitirse durante el proceso de selección del subconjunto de variables. Mientras que, en la extracción de variables, el tamaño del espacio de variables a menudo se puede disminuir sin perder mucha información del espacio de variables original. La elección entre la extracción de variables y los métodos de selección de variables depende del tipo de datos específico del dominio de aplicación.

Los datos de alta dimensionalidad contienen variables que pueden ser irrelevantes, engañosas o redundantes, lo que aumenta el tamaño del espacio de búsqueda, y provoca dificultades para procesar los datos por lo que no contribuye al proceso de aprendizaje.

La selección de subconjuntos de variables es el proceso de seleccionar las mejores variables entre todas las variables que son útiles para discriminar clases. Los algoritmos de selección de variables se pueden caracterizar por organización de búsqueda: exponencial, secuencial o aleatoria; generación de sucesores, se pueden considerar cinco operadores diferentes para generar sucesores que son hacia adelante, hacia atrás, compuesto, ponderado y aleatorio; medida de evaluación: la evaluación de sucesores se puede medir a través de la probabilidad de error, divergencia, dependencia, distancia entre clases, información o evaluación de incertidumbre y consistencia.

Los métodos de selección de variables se pueden distinguir en tres categorías: filtros, envoltorios y métodos híbridos. Los métodos de envoltura funcionan mejor que los métodos de filtro porque el proceso de selección de variables está optimizado para que se use el clasificador, pero su uso resulta oneroso para un gran espacio de variables debido al alto costo computacional y cada conjunto de variables debe evaluarse con el clasificador adecuado que finalmente hace que el proceso de selección de variables sea lento.

Los métodos de filtro tienen un bajo costo computacional y son más rápidos, pero con una confiabilidad menor en la clasificación en comparación con los métodos envolventes y son más adecuados para conjuntos de datos de alta dimensión. Los métodos híbridos se han desarrollado recientemente y utilizan las ventajas de los enfoques de filtros y envoltorios.

Un enfoque híbrido utiliza tanto una prueba independiente como una función de evaluación del rendimiento del subconjunto de variables (S. B. Kotsiantis & Pintelas, 2004). Un algoritmo de selección tiene como objetivo identificar las variables más importantes de acuerdo con una definición de relevancia (Ladha & Deepa, 2011).

En este trabajo se compararán ambas técnicas, con el fin de aplicar la que mejores resultados aporte, esto se evaluará por medio la precisión alcanzada.

Los algoritmos que se usarán para selección de variables son regresión logística y eliminación recursiva de variables, respectivamente. A continuación, se explicarán las bases teóricas del funcionamiento de cada uno de ellos.

3.2.1.1 Selección de variables a través de Regresión Logística simple

Para seleccionar las variables que están fuertemente asociadas con la clase minoritaria se aplica el modelo de regresión logística simple con regularización Ridge. La regresión logística permite cuantificar la probabilidad de pertenecer a la clase positiva (clase minoritaria) con respecto al incremento o decremento del valor de un atributo. Permite modelar la probabilidad de pertenecer a la clase minoritaria (P) como la función logística de la combinación lineal de los coeficientes del modelo $(\beta 0 + \beta 1)$ y cada uno de los atributos seleccionados (X). Los coeficientes del modelo se estiman a través de la función de máxima verosimilitud.

$$P = \frac{1}{1 + e^{-(\beta 0 + \beta 1 X)}} \qquad (6)$$

El objetivo primordial que resuelve esta técnica es el de modelar cómo influye en la probabilidad de aparición de un evento (dicotómico), la presencia o no de diversos factores de riesgo (atributos). El modelo de regresión logística simple es un modelo lineal generalizado que consta de tres componentes: un componente aleatorio Y (clase de los ejemplos) cuyo valor es 1 si ocurre el evento y 0 si no ocurre, y donde la probabilidad de que ocurra el evento es p y la probabilidad de que no ocurra es 1-p; un componente sistemático que es el conjunto de factores de riesgo X (atributos), los cuales actúan como variables

pronóstico asociadas a la ocurrencia del evento y una función de enlace entre ambos, llamada función logística (Tibshirani, 2011).

$$f(z) = \frac{1}{1 + e^{(-z)}} \qquad (7)$$

Una forma de representar el modelo es a través de la ecuación $\ln\frac{p}{q} = \beta_0 + \beta_1 X$, por lo que si el factor X es un atributo discreto binario que sólo puede tomar dos valores binarios, como ser o no consumidor de alcohol. Cuando el valor de X = 0 el modelo queda $ln\left(\frac{p}{q} \middle| X = 0\right) = \beta_0)$, de tal forma que β_0 es el logaritmo del odds (cociente de la probabilidad de que ocurra y no ocurra el evento) cuando la variable pronóstica no está presente, mientras que para el valor $X = 1$ queda $\ln\left(\frac{p}{q}\middle| X = 1\right) = \beta_0 + \beta_1$, por lo tanto, $\beta_1 = ln\left(\frac{p}{q}\middle| X = 1\right) - \beta_0$, sustituyendo β_0, la expresión finalmente se muestra como:

$$\beta_1 = \ln\left(\frac{\frac{p}{q} \mid X = 1}{\frac{p}{q} \mid X = 0}\right) \qquad (8)$$

Donde β_1 es el logaritmo del cociente de *odds* para los dos valores de X, esto es el logaritmo de OR.

Esto permite cuantificar la probabilidad de pertenecer a una clase. La regularización consiste en añadir una penalización a la función de costo. Esta penalización produce modelos más simples que generalizan mejor. Los métodos de regularización implican ajustar el modelo lineal incluyendo todas las variables disponibles usando una técnica que restringe o regulariza los estimadores de los coeficientes reduciéndolos hacia el cero. Las regularizaciones más usadas en aprendizaje automático son Lasso (también conocida como L1), Ridge (conocida también como L2) (Ladha & Deepa, 2011; Tibshirani, 1996).

Para este trabajo se hicieron pruebas con ambos métodos de regularización. Se seleccionó Lasso debido a que en la regularización Lasso la complejidad C se mide como la media del valor absoluto de los coeficientes del modelo.

$$C = \frac{1}{N}\sum_{j=1}^{N} |\omega_i| \qquad (9)$$

El error cuadrático medio es el criterio de evaluación más usado para problemas de regresión. Se calcula el cuadrado, con el fin de que sea siempre positivo. Para calcular el error medio se suman todos los errores y se dividen por el número total de puntos.

$$J = \frac{1}{M}\sum_{i=1}^{M}(real_i - estimado_i)^2 \qquad (10)$$

Al regularizar con Lasso, el error cuadrático medio se obtiene al sumar la complejidad al error cuadrático medio.

$$J = \frac{1}{M}\sum_{i=1}^{M}(real_i - estimado_i)^2 + \propto \frac{1}{N}\sum_{j=1}^{N}|\omega_i| \qquad (11)$$

Lasso es de utilidad cuando se sospecha que varios de los atributos de entrada son irrelevantes. Al usar la regularización Lasso, se fomenta que la solución sea poco densa. Es decir, que algunos de los coeficientes valgan 0. Esto puede ser útil para descubrir cuáles de los atributos de entrada son relevantes y, en general, para obtener un modelo que generalice mejor. Lasso funciona mejor cuando los atributos no están muy correlacionados entre ellos.

Ridge es útil cuando varias de las variables de entrada están correlacionadas entre ellos. Ridge hace que los coeficientes terminen siendo más pequeños, esta disminución de los coeficientes minimiza el efecto de la correlación entre los atributos de entrada y hace que el modelo generalice mejor. Ridge funciona mejor cuando la mayoría de los atributos son relevantes. En la regularización Ridge, también llamada L2, la complejidad *C* se mide como la media del cuadrado de los coeficientes del modelo.

$$C = \frac{1}{2N}\sum_{j=1}^{N}\omega_i^2 \qquad (12)$$

Al regularizar, al error cuadrático medio se le añade un término que penaliza la complejidad del modelo, en donde C es la medida de la complejidad del modelo. El hiper parámetro α indica la importancia de la simplicidad del modelo en relación con su rendimiento.

$$J = MSE + \propto \cdot C \qquad (13)$$

La regresión se hace por medio de validación cruzada con 10 particiones, en donde se divide en forma aleatoria el conjunto de datos en k subconjuntos disjuntos de tamaño similar, entonces se aprende una hipótesis utilizando el conjunto formado por la unión de k-1 subconjuntos (conjunto de entrenamiento), mientras que el subconjunto restante (conjunto de prueba) se usa para calcular el error parcial de la muestra.

El procedimiento se repite k veces, usando siempre un conjunto de prueba diferente (sin repetición). Esta forma de seleccionar los datos no garantiza que se mantenga el desequilibrio de clases inicial. Por esta razón se hará una validación cruzada estratificada, dividiendo previamente los datos por clase, para que a partir de ellos se construya cada uno de los k conjuntos de entrenamiento y de prueba, tratando de mantener la distribución de clases original (Dasgupta & Sun, 2011).

Cualquier característica que tenga coeficientes de regresión distintos de cero es 'seleccionada' por el algoritmo LASSO. Estos enfoques tienden a ser entre filtros y "*wrappers*" en términos de complejidad computacional (Tibshirani, 1996, 2011).

3.2.1.2 Eliminación recursiva de variables

La eliminación recursiva de variables (RFE, por sus siglas en inglés *Recursive Feature Elimination*) un método de selección de variables que se ajusta a un modelo y elimina la variable más débil hasta que se alcanza el número especificado de variables.

Las variables se clasifican según los atributos del modelo, y al eliminar de forma recursiva un pequeño número de variables por ciclo, RFE intenta eliminar las dependencias y la colinealidad que puedan existir en el modelo. RFE requiere un número específico de variables para mantener, sin embargo, a menudo no se sabe de antemano cuántas variables son válidas. Para encontrar el número óptimo de variables, se utiliza la validación cruzada con RFE para calificar diferentes subconjuntos de variables y seleccionar la mejor colección de variables. El método RFE utiliza *Random Forest* (RF, por sus siglas en inglés *Random Forest*) para medir la calidad de cada combinación de variables.

El algoritmo *Random Forest* (RF) genera conjuntos de árboles y los hace crecer. Frecuentemente se generan vectores aleatorios que gobiernan el crecimiento de cada árbol

en el conjunto. Un primer ejemplo es *Bagging*, donde se hace crecer una selección aleatoria (sin reemplazo) de los ejemplos en el conjunto de entrenamiento (Dietterich, 2000).

El algoritmo RF genera un vector aleatorio, Θ_k, independiente de los vectores $\Theta_1 \dots \Theta_{k-1}$ aleatorios anteriores con la misma distribución, se genera un árbol usando el conjunto de entrenamiento Θ_k lo que da como resultado un $h(\boldsymbol{x}, \Theta_k)$ clasificador donde **x** es un vector de entrada. Después de haber generado un gran número de árboles, se vota por la clase más popular.

Dado un conjunto de clasificadores $h_1(\boldsymbol{x})$, $h_2(\boldsymbol{x}) \dots h_k(\boldsymbol{x})$, con un conjunto de entrenamiento generado aleatoriamente de la distribución del vector aleatorio **X,**Y, se define la función marginal como:

$$mg(\boldsymbol{X}, Y) = av_k I(h_k(\boldsymbol{X}) = Y) - \max_{j \neq Y} av_k I(h_k(\boldsymbol{X}) = j) \cdot \qquad (14)$$

En donde (·) es el identificador de función. La función marginal mide en qué medida el número medio de votos en **X**,**Y** para la clase correcta excede el voto promedio para cualquier otra clase. Cuanto mayor sea el margen, más confianza se tendrá en la clasificación. El error de generalización está dado por:

$$PE^* = P_{\boldsymbol{X},Y}(mg(\boldsymbol{X}, Y) < 0) \qquad (15)$$

En donde **X**,**Y** indican que la probabilidad es sobre el espacio **X**,**Y**

Conforme el número de árboles aumenta, para casi todas las secuencias $\Theta_1 \dots$PE* convergen a:

$$mg(\boldsymbol{X}, Y) = av_k I(h_k(\boldsymbol{X}) = Y) - \max_{j \neq Y} av_k I(h_k(\boldsymbol{X}) = j) \quad < 0 \qquad (16)$$

Esto explica por qué RF no sobre ajusta al agregar más árboles, pero produce un valor limitante de la generalización del error (Breiman, 2001).

El método funciona de la siguiente forma: en primer lugar, el algoritmo RFE ajusta el modelo a todas las variables, cada variable se clasifica según la importancia del modelo. Luego, el algoritmo RFE comienza a crear modelos utilizando las variables $S\_i, i = 1 \dots S$.

Entonces intenta todas las combinaciones posibles y mantiene en una lista la combinación de variables y su rendimiento. Para cada iteración, todas las variables se clasifican nuevamente. Al final de la ejecución del algoritmo, se realiza una lista de clasificación utilizando los resultados de todas las iteraciones. Finalmente, se selecciona la combinación con la mayor precisión.

La elección de la técnica más adecuada para la selección de variables debe ser elegida con base en las variables de los datos del dominio estudiado. En este trabajo se aplican ambas y se hace una comparación de los resultados obtenidos en los conjuntos de datos de DMRE y Preeclampsia.

3.3 Balanceo de las clases

Los clasificadores simbólicos asignan la clase a una instancia por medio de condiciones excluyentes y exhaustivas que dividen el espacio de variables (Tan et al., 2018). Esto evidencia que es necesario encontrar una proporción adecuada entre exactitud y eficiencia que permita elegir la mejor partición de los datos. En este punto es necesario resolver el problema del desbalanceo de las clases que por sí sólo provoca el mal desempeño de los algoritmos de clasificación.

El problema de desequilibrio de clase ocurre cuando hay una gran diferencia entre el número de clase mayoritaria y la clase minoritaria y principalmente en clases con valores binarios (Batista et al., 2004). La disparidad causada en los valores de la clase objetivo podría tener un impacto extremadamente negativo en el rendimiento de los algoritmos de aprendizaje automático (Menardi & Torelli, 2014). La mayoría de las veces conduciría a una clasificación falsa y el resultado de la predicción se sobre ajustará porque el modelo no atenúa el sesgo para la clase mayoritaria o no se realiza correctamente debido a los muy pocos casos de clase positiva. En la práctica se ha demostrado que, en muchos casos, se logra mejor desempeño usando conjuntos de datos balanceados, para lograrlo existen métodos llamados métodos de muestreo.

Un método que ha demostrado funcionar bien para resolver el problema del desequilibrio de clase es "*Synthetic Minority Oversampling Technique*" (SMOTE) (Blagus & Lusa, 2013). Es un método de sobre muestreo que genera objetos sintéticos en puntos

aleatorios entre objetos de la clase minoritaria y alguno de sus k-vecinos más cercanos elegido al azar. SMOTE trabaja de la siguiente manera:

Calcula la cantidad de objetos sintéticos a generar (n), de acuerdo con un parámetro $N \in \mathbb{R}$, que es la proporción a sobre muestrear en la clase minoritaria (m). ($n = |m| * N$, donde $|m|$ es la cantidad de objetos en la clase minoritaria).

- Si $N < 1$, la cantidad de objetos a generar es menor que la cantidad de objetos en la clase minoritaria, por lo que se selecciona un subconjunto aleatorio, de tamaño n, de objetos de la clase minoritaria.
- Si $N = 1$, la cantidad de objetos a generar es igual a la cantidad de objetos de la clase minoritaria, por lo que se seleccionan todos los objetos de la clase minoritaria.
- Si $N > 1$, la cantidad de objetos a generar es mayor a la cantidad de objetos en la clase minoritaria, en este caso se seleccionan todos los objetos de la clase minoritaria tantas veces como el valor entero de N, además de un subconjunto aleatorio de objetos de la clase minoritaria en función de la diferencia entre N y el valor entero N.

Para cada objeto seleccionado de la clase minoritaria, se elige aleatoriamente uno de sus k vecinos más cercanos, donde k es un parámetro, y se siguen con los siguientes pasos:

- Se calcula la diferencia para cada atributo del objeto de la clase minoritaria y su vecino más cercano seleccionado.
- Se multiplica la diferencia por un valor aleatorio entre 0 y 1.
- Se suma el resultado de la multiplicación al objeto original de la clase minoritaria, generando de esta manera un nuevo objeto sintético (ver ecuación 14)

$$x_{nueva} = x_i + \left(x_i^k - x_i\right) \times \delta \qquad (14)$$

En donde x_i^k es uno de los vecinos más cercanos de x_i y $\delta \in [0,1]$ es un número aleatorio. De manera que la instancia generada es un punto a lo largo de del segmento de línea que une a x_i y al vecino más cercano elegido al azar x_i^k.

Esto permite controlar la posición final de la instancia artificial, que se puede ubicar en la misma posición de la instancia original, la instancia vecina seleccionada o entre ellas, dependiendo del valor generado aleatoriamente, ver Figura 3.4 . Esto aumenta la diversidad del conjunto de instancias artificiales, lo que permite una mejor explotación del espacio de decisión dado (Skryjomski & Krawczyk, 2017).

Figura 3.4 Algoritmo de sobre muestreo por medio de la técnica SMOTE

$Función\ SMOTE(T_{min}, N, k)$
$T \leftarrow [\ ]$
$Para\ i \leftarrow 1\ to\ nrenglones(T_{min})\quad hacer$
    $nn \leftarrow kNN(T_i, T_{min}, k)$
    $N_i \leftarrow [N/100]$
    $Mientras\ N_i! = 0\ hacer$
        $vecino \leftarrow selección - aleatorio(nn)$
        $gap \leftarrow numaleatorio(0,1)$
        $dif \leftarrow vecino - T_i$
        $sintético \leftarrow T_i + gap * dif$
        $T_{SMOTED} \leftarrow agregar(T_{SMOTED}, sintético)$
        $N_i \leftarrow N_i - 1$
    Fin
Fin

SMOTE tiene algunos inconvenientes como generar muchos ejemplos artificiales cuyas semillas son ejemplos con ruido; al generar un nuevo ejemplo, interpola entre dos ejemplos de la clase minoritaria, pero pueden existir muchos ejemplos cercanos o inclusive entre ellos de la clase mayoritaria, generando modelos incorrectos.

Con el paso del tiempo se han hecho propuestas para mejorar el desempeño de SMOTE. Entre ellas está cambiar la forma de medir las distancias entre los vecinos más cercanos; etiquetar los ejemplos de la clase minoritaria que estén cerca de los de la clase mayoritaria, para evitar generar ejemplos espurios; generación aleatoria de ejemplos; combinaciones con otros métodos como el sub muestro, filtrado de ruido, generación de reglas, entre otros.

Los métodos mencionados eligen de forma aleatoria los ejemplos o los vecinos más cercanos para generar los ejemplos sintéticos. Esto provoca el problema de producir resultados distintos cada vez que se aplican estos métodos a un mismo conjunto de datos. Al funcionar de esta manera, existe la posibilidad de que los objetos sintéticos generados sesguen su posición hacia un objeto particular sin razón, lo que puede provocar que los resultados sean deficientes en la clasificación de los datos. Para solucionar esto se recurre a técnicas como Tomek (Batista et al., 2004), que eliminan los datos espurios.

El objetivo de este algoritmo es aclarar la frontera entre las clases minoritarias y mayoritarias, haciendo que las regiones minoritarias sean notablemente distintas a las de la clase mayoritaria para favorecer el trabajo de clasificación.

3.3.1 **Selección del ensamble adecuado para los conjuntos de datos**

Se aplicaron dos ensambles diferentes para Boosting y la misma cantidad para Bagging. Se probaron cinco clasificadores, que se apilaron por medio de GLM, con el fin de contar con elementos suficientes para decidir cuál es el ensamble que resultó más conveniente para el caso de estudio.

Una vez que se apliquaron los ensambles diseñados bajo los tres enfoques mencionados, se evaluó cuál fue en ensamble indicado para cada conjunto de datos en particular, con base en las métricas *Accuracy* y Kappa. En este punto es necesario establecer la diferencia entre Kappa y *Accuracy*, con el fin de decidir cuál será la métrica que se usará para decidir qué clasificador es el más adecuado para el conjunto de datos en estudio.

Accuracy es el porcentaje de clasificaciones correctas de todas las instancias. Es una descripción de errores sistemáticos, por lo tanto, es una descripción de errores aleatorios.

Kappa es una métrica que se usa para evaluar a un clasificador o a varios clasificadores entre sí e indica cómo se desempeñó el clasificador en todas las instancias. *Accuracy* puede alterarse si la distribución de clases está sesgada, lo que sucede en los conjuntos de datos con clases desbalanceadas. Una gran ventaja de Kappa es que la medición realizada en un modelo es directamente comparable con el de cualquier otro modelo utilizado para la misma tarea de clasificación.

En este trabajo se compara el desempeño de varios clasificadores, lo que hace que Kappa sea una métrica adecuada.

3.4 Interpretabilidad

Los resultados de los pasos previos deben mostrarse de forma clara y detallada, con el fin de facilitar que los usuarios los utilicen como apoyo en la toma de decisiones.

La presentación de resultados de manera que resulten interpretables se hace tomando en cuenta los tres niveles de interpretabilidad:

i. Interpretabilidad de los datos,
ii. Interpretabilidad de los algoritmos aplicados para llegar a las predicciones.
iii. Interpretabilidad de las predicciones

El primero se refiere a qué datos se usaron para entrenar el modelo y por qué se usaron. Los datos se caracterizan a través de la descripción de su origen y condiciones de recolección, del tamaño de la muestra, el número y tipo de variables presentes, la frecuencia de los valores presentes en cada variable y la cantidad de ejemplos para cada clase.

Esto se detalla de la siguiente forma: listado de las variables presentes en el conjunto de datos; los tipos de datos a que pertenecen; sus frecuencias y rangos, para las variables categóricas y numéricas respectivamente; el número total de instancias y el número de las pertenecientes a cada clase.

El segundo nivel se refiere a la forma en que trabajan los ensambles de clasificadores. Los algoritmos utilizados para construir los ensambles de clasificadores se presentan por medio de interpretabilidad *post-hoc*. Esto es, se extrae información del modelo que ya ha sido aprendido. Se parte de que la mejor explicación de un modelo simple es el modelo mismo; se representa perfectamente a sí mismo y es fácil de entender (Rodríguez Torres, 2017). Este enfoque se limita a las familias de modelos con menor complejidad, como modelos lineales, árboles de decisión y reglas.

Debido a que los modelos complejos se consideran opacos y su complejidad impide a los usuarios rastrear la lógica detrás de las predicciones, se recurre a los modelos menos complejos. Esto se hace por la sencillez en su interpretación sin pérdida significativa en el desempeño con respecto a los modelos más complejos.

Los algoritmos utilizados para construir los ensambles de clasificación de los conjuntos de datos se enlistan y se explica de manera sucinta el funcionamiento de cada uno de ellos en el conjunto de datos particular.

El tercer nivel se refiere a la explicación de cómo se llegó a las predicciones propuestas, esto implica explicar qué variables o características se usaron para una predicción. La explicación se presenta de manera global y local.

La forma en que se llega a las predicciones se explica de manera global por medio de árboles de decisión y reglas de inferencia. A los expertos en el dominio médico puede resultarles de interés la forma en que se encuentran los factores de riesgo para una enfermedad para cierta población. Esto es, de la representación global puede extraerse conocimiento aplicable a grupos étnicos definidos en el diseño del estudio.

La interpretabilidad global se puede representar visualmente a través de árboles de decisión, histogramas, nomogramas, entre otros. También se puede explicar la forma de llegar a las predicciones a través de las reglas de inferencia generadas a partir de los árboles de decisión. Las reglas de inferencia pueden resultar un lugar intermedio entre los modelos opacos y los que se explican de manera sencilla.

En lógica computacional, las bases de conocimiento se conforman a partir de los datos y el conocimiento generado. Para representar el conocimiento generado, se hace uso de las reglas de inferencia. Las reglas de inferencia son proposiciones lógicas que relacionan a dos o más objetos del dominio e incluye dos partes, la premisa y la conclusión, que se escribe: "Si premisa, entonces conclusión". Cada una de estas partes es una expresión lógica con una o más afirmaciones variable-valor conectadas mediante operadores lógicos (Caparrini & de J. Pérez Jiménez, 2002).

Las reglas de inferencia tienen la ventaja de representar de forma natural el conocimiento explícito de los expertos, que generalmente explican el procedimiento de resolución de problemas por medio de expresiones del tipo "Si *problemática*, entonces *lo resolvería de tal manera…*". Adicionalmente, las reglas tienen estructura uniforme y cada regla es una pieza de conocimiento independiente de las demás.

Con el fin de mejorar la interpretabilidad de las predicciones, en este trabajo se muestran las reglas de inferencia generadas a partir de los árboles de decisión como motores de inferencia, además de presentarse los árboles de decisión, histogramas y nomogramas.

De manera local, se explica bajo el enfoque agnóstico[2] a los modelos de clasificación, de acuerdo con el tipo de usuario final al que están destinados los resultados. En el dominio médico es importante mostrar con claridad la forma en que se llega a los resultados en lo global, lo que permite la construcción del conocimiento a partir de los resultados; y de manera local para conocer los mecanismos usados para instancias particulares. Lo que implica mostrar al usuario médico los resultados obtenidos para los individuos seleccionados, de manera particular.

A pesar de que un modelo interpretable puede no explicar un modelo complejo globalmente, puede resultar eficiente para explicarlo en la vecindad de una instancia, con fidelidad al modelo de clasificación. Lo que hace accesible la forma de llegar a los resultados a los usuarios que no necesariamente son expertos en los modelos de aprendizaje automático. Las explicaciones se representan gráficamente en términos de las variables de estudio presentes en el conjunto de datos.

La interpretabilidad agnóstica al modelo es un componente clave que permite que los resultados sean más confiables y útiles para los usuarios del dominio médico. Esto significa que las explicaciones deben ser válidas para cualquier modelo y no deben hacerse suposiciones sobre ningún modelo al proporcionar explicaciones. La representación interpretable a nivel local es una combinación ponderada de las variables del conjunto de datos.

En este capítulo se presentó la metodología que se aplicará para determinar los factores de riesgo asociados a enfermedades complejas mediante el enfoque de aprendizaje automático.

A continuación, se expondrán los dos escenarios de aplicación de la metodología presentada: Degeneración Macular Relacionada con la Edad y Preeclampsia.

[2] Un modelo agnóstico es independiente de los algoritmos aplicados para la clasificación.

Capítulo 4 Presentación y análisis de resultados

En esta sección se presentan los resultados obtenidos en esta investigación en dos escenarios de aplicación. Para mostrar la pertinencia de la metodología propuesta se eligieron dos enfermedades complejas, el estudio de ellas bajo el enfoque de aprendizaje automático ilustra los principales retos propios del dominio de cuidados de la salud, estas son DMRE y Preeclampsia.

Los retos por afrontar en este dominio son principalmente la escasez de datos útiles, el desbalanceo en los conjuntos de datos, la necesidad de interpretabilidad y transparencia en los resultados obtenidos.

Ambos escenarios de aplicación están enfocados a enfermedades complejas presentes en una creciente proporción de la población. El primer escenario de aplicación es el de la DMRE, que resultó de particular interés porque la enfermedad ha sido ampliamente estudiada en etnias diferentes a la mexicana.

Con el fin de contribuir al estudio de la enfermedad en mexicanos, se obtuvieron los datos sociodemográficos, clínicos y ADN en el Instituto de Oftalmología Conde de Valenciana ubicado en la Ciudad de México. La base de datos construida con estos datos fue de interés por el reto que significó comenzar con la obtención de muestras de ADN para genotiparlas en el laboratorio de Biología Molecular de la Universidad Panamericana, la selección de los datos útiles para los casos y controles incluido en este trabajo, aunque la base de datos obtenida no tenía un desbalanceo significativo, se mejoró el balanceo. Un desafío de este escenario de aplicación fue la presentación interpretable para los especialistas en el área de oftalmología.

Para el escenario de aplicación enfocado en preeclampsia, se tomó como base de datos de referencia de un estudio realizado por el gobierno de Estados Unidos de Norteamérica, en particular en Trenton, Nueva Jersey. Los datos disponibles en el conjunto de datos fueron sociodemográficos y ambientales de las mujeres embarazadas. Esta base de datos resultó de interés por la cantidad de mujeres que presentan esta enfermedad en el mundo, por las graves consecuencias de esta y sobre todo por su alto grado de desbalanceo (19%) y la necesidad de la presentación interpretable de los resultados.

Los resultados de cada uno de los escenarios de aplicación se obtuvieron aplicando la metodología planteada previamente. Estos se muestran a continuación de forma detallada para cada uno de ellos.

La metodología propuesta indica que una vez que se ha construido la base de datos se busquen los datos faltantes, para reemplazarlos por la media o la mediana, para variables numéricas y variables categóricas, respectivamente. A continuación, se procede a seleccionar las variables más relevantes para la clasificación por medio de dos métodos, que se comparan con el fin de seleccionar el que resulte más eficiente. Se mostrarán los conjuntos de datos balanceados obtenidos por medio de sobre muestro aleatorio y sobre muestreo de la clase minoritaria con sub muestro de la clase mayoritaria por medio de SMOTE. Las variables categóricas se convierten en vectores binarios con el fin de evitar sesgos. Posteriormente, se miden las distancias entre las instancias de diferentes clases para identificar a los pares que estén muy cercanos entre sí. Una vez identificados, de cada par se elimina la que pertenezca a la clase mayoritaria. Se construyen tres tipos de ensambles, cuyo desempeño se compara a través de las métricas Accuracy y Kappa.

Los resultados de la clasificación se evalúan a través de las curvas ROC obtenidas para cada conjunto de datos y se muestran los resultados de las métricas de evaluación: AUC (área bajo la curva), CA (*Accuracy*), F1, Precisión, Sensibilidad. También se presentan gráficas comparativas para el área bajo la curva (AUC) de los clasificadores probados.

En seguida, se presentan los resultados de manera globalmente interpretable por medio de árboles de decisión, nomogramas y reglas de inferencia. Esto pretende dar los elementos suficientes para que se comprenda como se determinaron los factores de riesgo para los escenarios de aplicación.

Así mismo, los resultados se presentan de manera que resulten localmente interpretables. Con el fin de que los expertos del ámbito médico cuenten con información de los individuos de su interés.

4.1 Primer escenario de aplicación: Degeneración Macular Relacionada con la Edad

El escenario de aplicación Degeneración Macular Relacionada con la Edad (DMRE) se basó en un estudio de casos y controles en una población de 256 mexicanos no relacionados, con edad mayor a 60 años, con 119 casos de DMRE y 137 controles.

4.1.1 Construcción de la base de datos

La selección de las variantes genéticas se basó en el estudio de asociación del genoma amplio (GWAS), que analiza la contribución a la DMRE de las variaciones genéticas que son comunes en una población (Hindorff et al., 2014). También se consideraron estudios más astringentes para reducir las regiones del genoma que confieren más susceptibilidad a la enfermedad (Chakravarthy et al., 2010). Con base en esos estudios, se seleccionaron dos polimorfismos del gen CFH, y uno en HTRA1 que son genes identificados por generar riesgo de DMRE entre varias poblaciones.

De esta manera, los polimorfismos de nucleótido único (SNP) rs1329428, rs203687 en CFH y rs11200638 en HTRA1 se genotiparon a partir de muestras de ADN de cada uno de los sujetos bajo estudio bajo el procedimiento que se detalla a continuación.

Se recogieron muestras de sangre periférica en tubos con EDTA; Se extrajo y purificó el ADN genómico de los leucocitos utilizando el kit de sangre completa PureGene (QIAGEN, Germantown, MD, EE. UU.) siguiendo las especificaciones del fabricante; La concentración y pureza del ADN se cuantificaron utilizando un espectrofotómetro Multiskan TM GO (Thermo Fisher Scientific Inc., Wilmington, DE, EE. UU.). Además, evaluamos la integridad del ADN con gel de agarosa al 0.8% (Thermo Fisher Scientific Inc.) teñido con bromuro de etidio.

La discriminación alélica se realizó utilizando sondas TaqMan prediseñadas (Thermo Fisher Scientific Inc., en reacción en cadena de la polimerasa en tiempo real (RT-PCR) (Piko Real, Thermo Fisher Scientific Inc.). Todas las amplificaciones de PCR incluyeron 6.5 µL de Maxima Probe qPCR Master Mix 2X (Thermo Scientific), 0.025 µL de cebadores, 20 × sondas (Thermo Fisher Scientific), 20 ng de ADN total y agua libre de nucleasas, en un volumen final de 10 µL. Las condiciones térmicas de la PCR fueron 10 min a 95 ◦C, seguidas

de 40 ciclos a 92 ◦C durante 15 segundos, extendido a 60 ◦C durante 1 min. Se empleó análisis de curva de fusión de RT-PCR convencional para la asignación de genotipos utilizando el sistema de PCR en tiempo real (Thermo Fisher Scientific Inc., Waltham, MA, EE. UU.). Todos los ensayos se llevaron a cabo por duplicado.

Para realizar este trabajo se contó con la aprobación del Hospital Conde de Valenciana (IRB CEI-2014- / 02/01). Las muestras se obtuvieron siguiendo estrictamente los principios de la Declaración de Helsinki (World Medical Associationdeclaration of Helsinki, 2014).

Las personas afectadas tuvieron un diagnóstico clínico de DMRE con atrofia geográfica o enfermedad neovascular. El diagnóstico y la estratificación siguieron las actuales directrices clínicas de la Asociación Americana de Oftalmología (Goldberg et al., 1988). Los controles fueron participantes sin ninguna evidencia de DMRE avanzado. Las variables demográficas, así como los factores de riesgo conocidos para la DMRE (edad, sexo, estado de tabaquismo y comorbilidades adicionales hipertensión, diabetes y dislipidemia), se obtuvieron de los registros médicos electrónicos (ver Tabla 4.1).

Tabla 4.1 Descripción de la muestra para DMRE

Características	**N**	**%**	**Sí (N =119)**	**%**	**No (N=137)**	**%**
Demográficas						
Edad, Media ± Desviación std	74.0 ± 8.2		75.1 ± 8.3		73.1 ± 8.0	
Sexo:						
Mujeres	164	64.06	64	25.00	100	39.06
Hombres	92	35.94	55	21.48	37	14.45
Comorbilidades críticas						
Hipertensión	130	50.78	93	36.33	37	14.45
Consumo de tabaco	52	20.31	52	20.31	19	7.42
Consumo de Alcohol	23	8.98	16	6.25	7	2.73
Diabetes tipo 2	57	22.27	28	10.94	29	11.33
Dislipidemia	12	4.69	9	3.52	3	1.17
Comorbilidades oculares						
Retinopatía diabética	9	3.52	2	0.78	7	2.73
Presbiopía	60	23.44	28	10.94	32	12.50

Características	N	%	Sí (N =119)	%	No (N=137)	%
Glaucoma	56	21.88	28	10.94	28	10.94
Historia de cirugía oftálmica	91	35.55	31	12.11	60	23.44
Distribución de genotipos						
CFH (rs1329428)						
CC	25	9.77	11	4.30	14	5.47
CT	96	37.50	48	18.75	48	18.75
TT	39	15.23	17	6.64	22	8.59
CFH (rs203687)						
CC	25	9.77	11	4.30	14	5.47
CT	223	87.11	105	41.02	118	46.09
TT	8	3.13	3	1.17	5	1.95
HTRA1 (rs11200638)						
GG	79	30.86	44	17.19	35	13.67
AG	160	62.50	64	25.00	96	37.50
AA	17	6.64	11	4.30	6	2.34

Fuente: Elaboración propia.

4.1.2 **Selección de las variables más relevantes para DMRE**

El primer paso para el manejo adecuado de los conjuntos de datos fue la búsqueda de datos faltantes. En el caso del conjunto de datos DMRE los ejemplos con datos faltantes se sustituyeron por la media o la mediana, si la variable era numérica o categórica, respectivamente. La cantidad de faltantes fue de 1%, por lo que no significó pérdida de información.

Para seleccionar las variables de mayor relevancia para la clasificación del conjunto de datos de DMRE, se aplicó RFE con *Random Forest.*

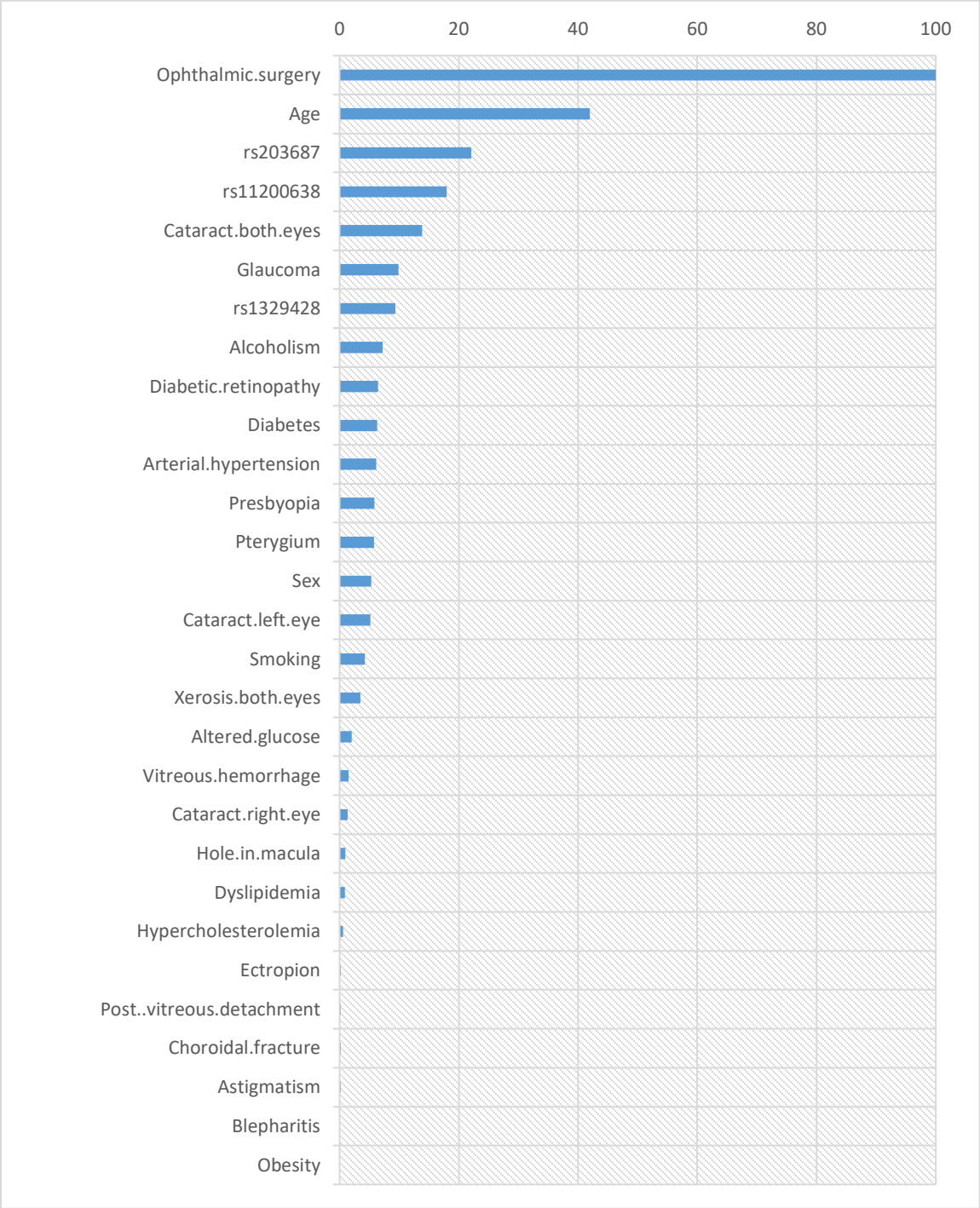
Cada variable se clasificó de acuerdo con su importancia para el modelo. Los modelos generados se evaluaron con la métrica "*Accuracy*" para determinar el poder predictivo del conjunto de variables en el proceso de clasificación.

El algoritmo RFE ajusta el modelo a las 29 variables presentes en el conjunto de datos. Posteriormente, el algoritmo crea modelos utilizando S_i variables, con $i = 1 \dots S$.

El algoritmo RFE probó todas las combinaciones posibles y las almacenó en una lista de combinación de variables y su rendimiento.

Para cada iteración, todas las variables se clasifican nuevamente. Al final de la ejecución del algoritmo, se realiza una lista de clasificación utilizando los resultados de todas las iteraciones (ver Tabla 4.2), eso explica por qué el orden de las variables es diferente en la Figura 4.1. Con base en los resultados de la Tabla 4.2, se selecciona la combinación con la mayor precisión.

Figura 4.1 Importancia de las variables para el conjunto de datos DMRE



Fuente: Elaboración propia.

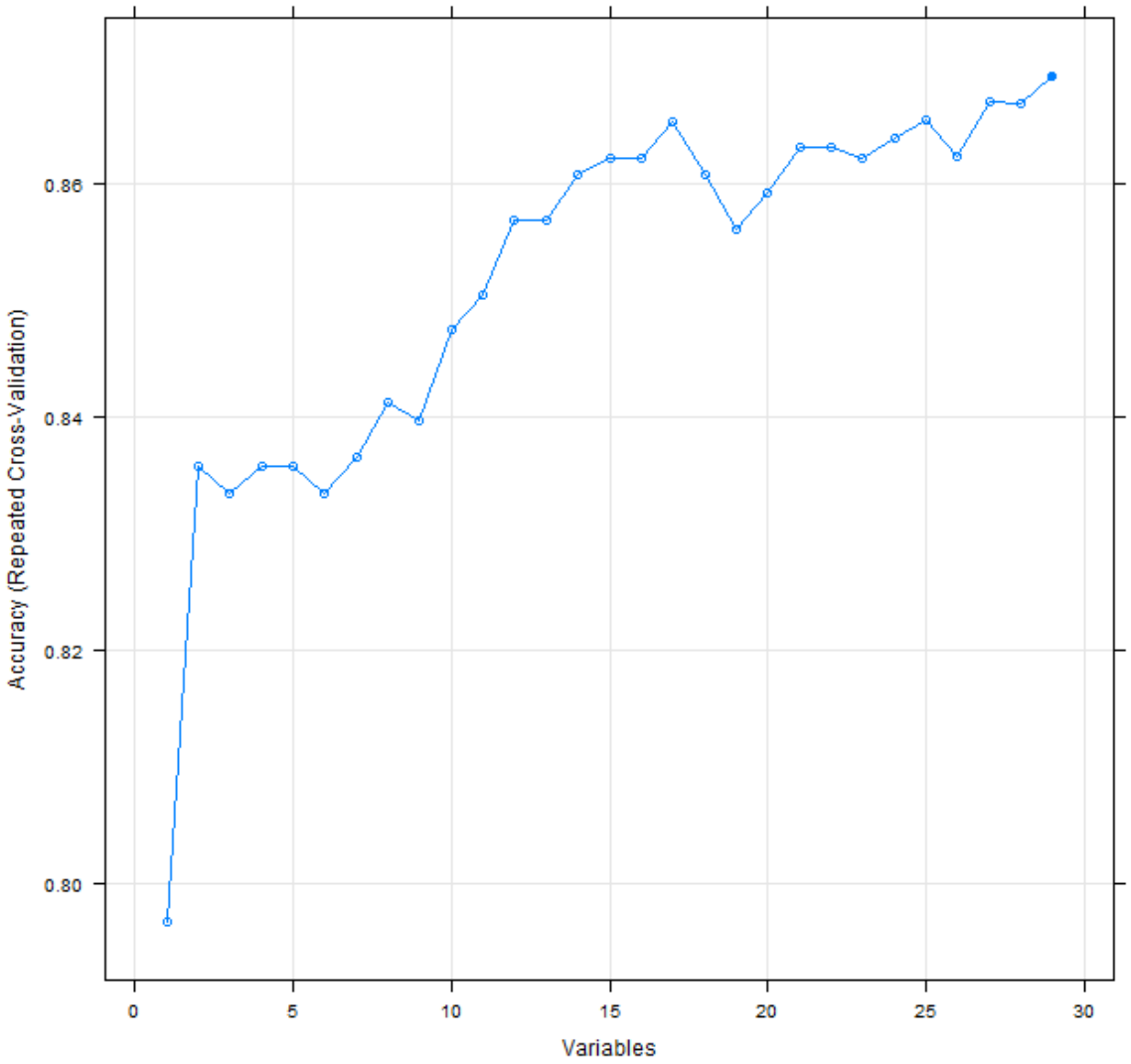
Tabla 4.2 Importancia de las variables y *Accuracy* combinada para el conjunto de datos

N°	Variable	Accuracy	Kappa	AccuracySD	KappaSD
1	Cirugías oftalmológicas	0.7969	0.6024	0.05913	0.1130
2	Polimorfismo rs203687	0.8242	0.6537	0.05837	0.1129
3	Cataratas en ambos ojos	0.8332	0.6708	0.06078	0.1183
4	Polimorfismo rs11200638	0.8245	0.653	0.06297	0.1233
5	Ingesta de alcohol	0.8258	0.6549	0.06180	0.1211
6	Agudeza visual en el ojo izquierdo	0.8184	0.6397	0.06295	0.1236
7	Agudeza visual en el ojo derecho	0.8188	0.6398	0.06689	0.1318
8	Glaucoma	0.8368	0.6755	0.07229	0.1432
9	Caratata en ojo izquierdo	0.8501	0.7012	0.06397	0.1271
10	Pterigión	0.861	0.7233	0.06167	0.1218
11	Retinopatía Diabética	0.8695	0.7405	0.05935	0.1171
12	Glucosa alterada	0.8683	0.7383	0.05875	0.1158
13	Diabetes	0.8699	0.7411	0.05917	0.1172
14	Hemorragia vitrea	0.8722	0.746	0.05805	0.1146
15	Edad	0.8691	0.7397	0.05900	0.1163
16	Obesidad	0.8656	0.7324	0.05705	0.1127
17	Sexo	0.8683	0.7377	0.05258	0.1044
18	Hipercholesterolemia	0.8707	0.7426	0.05774	0.1144
19	Xerosis ambos ojos	0.8715	0.744	0.05671	0.1125
20	Catarata ojo derecho	0.8696	0.7401	0.05762	0.1144
21	Presbiopia	0.8704	0.7418	0.05670	0.1127
22	Polimorfismo rs1329428	0.8715	0.7442	0.05557	0.1101
24	Astigmatismo	0.868	0.7374	0.05862	0.1161
25	Tabaquismo	0.8708	0.7425	0.05397	0.1072
26	Agujero en mácula	0.8708	0.7424	0.05478	0.1088
27	Bleparitis	0.8715	0.744	0.05447	0.1082
28	Fractura en vitreous vitero posterior	0.8747	0.7503	0.05358	0.1063
29	Fractura Coroidal	0.8738	0.7486	0.05617	0.1114
30	Dislipidemia	0.8743	0.7493	0.05606	0.1113
31	Ectropión	0.8758	0.7526	0.05625	0.1115

Fuente: Elaboración propia.

En la Figura 4.2 se muestra que, a partir de la combinación de cinco variables, los modelos no mejoran significativamente. Por lo que se decidió incluir sólo el conjunto de las primeras cinco de ellas: Cirugías oftalmológicas, Polimorfismo rs203687, Cataratas en ambos ojos, polimorfismo rs11200638 e Ingesta de alcohol.

Figura 4.2 Conjuntos de variables analizadas ordenadas por su *Accuracy* con RFE usando validación cruzada



Las variables seleccionadas como más relevantes permitirán reducir el tamaño del conjunto de datos y hacer más eficientes los pasos subsecuentes para continuar con su clasificación.

4.1.3 **Balanceo del conjunto de datos DMRE**

El conjunto de datos DMRE, tiene un desbalanceo moderado, con una tasa de desbalanceo es de 0.8686131 y la proporción entre los casos y controles es de 0.4648438 para los casos, y 0.5351562 para los controles.

Para el conjunto de datos DMRE, SMOTE con 1/1 aumentó la muestra de 238 instancias de la clase minoritaria y 238 de la clase mayoritaria, con tasa de desbalanceo (IR)

de 1.0, SMOTE con 2/1 aumentó la muestra de 357 instancias de la clase minoritaria 476 instancias de clase mayoritaria con tasa de desbalanceo (IR) de 0.75. La tasa de desbalanceo hace evidente que, en este caso, no es necesario el sobre muestro más allá de la proporción 1/1 (ver Tabla 4.3).

Tabla 4.3. Instancias sobre y sub muestreadas con de SMOTE para DMRE.

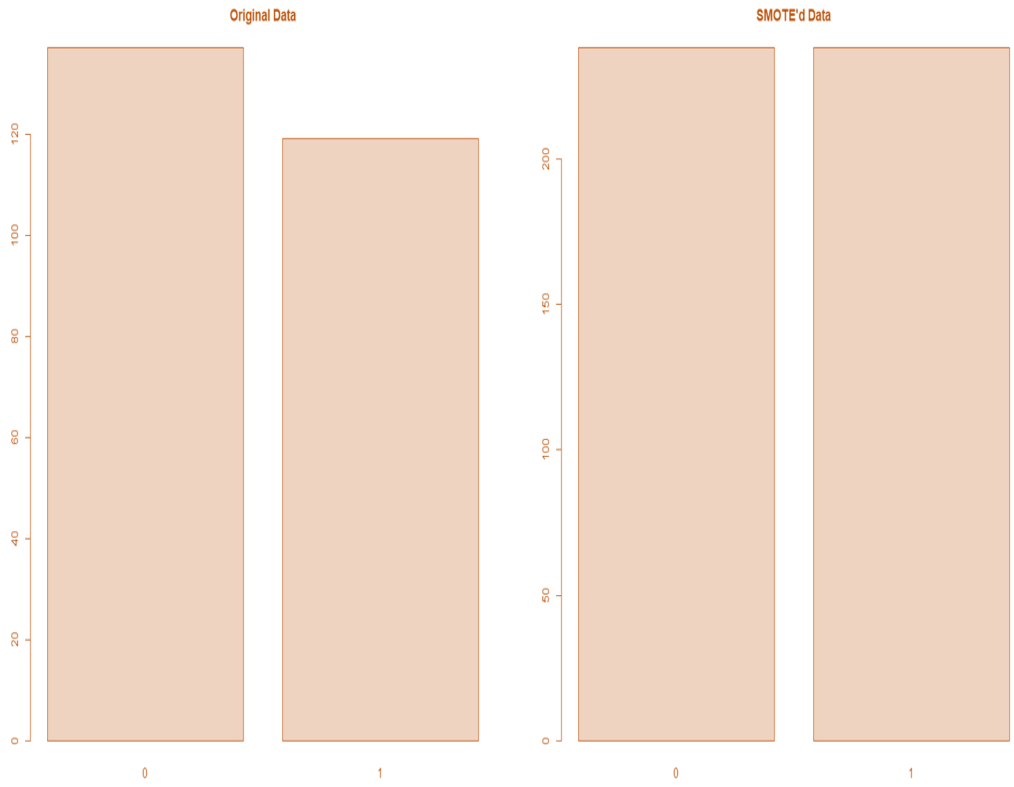
Proporción de sobre muestreo y submuestreo	Casos	Controles	Tasa de desequilibrio
ORIGINAL	137	119	0.8686131
1/1	**238**	**238**	**1**
2/1	476	357	0.75

Fuente: Elaboración propia.

En la Figura 4.3 Conjunto de datos DMRE antes y después de aplicar SMOTE con proporción de sobre muestreo 1/1 se aprecia que se logró mejorar el equilibrio entre clases,

aumentando la clase minoritaria y la clase mayoritaria hasta alcanzar el mismo número de instancias para cada una.

Figura 4.3 Conjunto de datos DMRE antes y después de aplicar SMOTE con proporción de sobre muestreo 1/1



Fuente: Elaboración propia

Una vez balanceado el conjunto de datos, se eliminarán los datos espurios con el fin de mejorar los resultados obtenidos durante los pasos subsecuentes.

4.1.4 **Eliminación de instancias espurias del conjunto de datos DMRE**

Las variables nominales, son usadas para nombrar o categorizar información. Este tipo de datos no están ordenados, incluso si se usan números para representarlos. Esto puede interferir con la clasificación pues el algoritmo puede asumir que las enumeraciones corresponden a algún tipo de jerarquía entre las instancias. Con el fin de evitar esto, se transformaron los datos para que cada categoría de la variable categórica en cuestión tenga un valor numérico.

La mayor parte de las variables del conjunto de datos DMRE son categóricas con valores 0 o 1 para las variables demográficas y de comorbilidades y C/C, C/T, T/T para los polimorfismos. Esto se realizó por medio del algoritmo de codificación *"One hot"*.

Una vez generados los vectores correspondientes, se utilizó el algoritmo *Tomek links* para eliminar las instancias de la clase mayoritaria que están muy cerca de instancias de la clase minoritaria. Para el conjunto de datos de DMRE se eliminaron 34 instancias, lo que representa el 14.29 % de los controles.

Ya que se han resuelto los problemas de desbalanceo de clases, el de las variables categóricas y el de los datos espurios, se prosigue con la clasificación de los conjuntos de datos.

4.1.5 **Ensambles de datos para clasificar el conjunto de datos DMRE**

Si bien, el conjunto de datos DMRE no tenía un desbalance severo entre sus dos clases, se balancearon los datos de manera que al momento de comenzar la tarea de clasificación el conjunto estuviera más equilibrado y el trabajo de los ensambles sea lo más eficiente posible. Para lograrlo, se usaron tres enfoques distintos de ensambles: *Boosting, Bagging y* Apilamiento (*Stack).*

Ensambles Boosting en el conjunto de datos DMRE

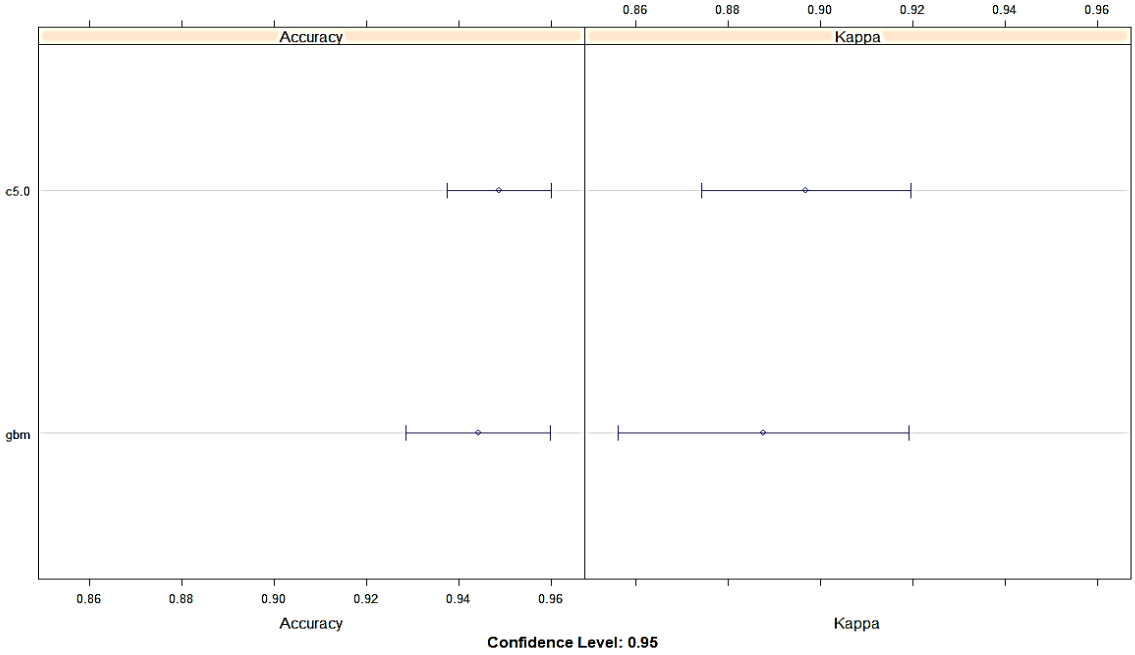
Los resultados de la clasificación por medio de ensambles *Boosting*, se presenta en la Tabla 4.4, en donde se evalúa el desempeño de los ensambles C5.0 y GBM. El ensamble C5.0 obtiene un valor ligeramente mayor que GBM con las métricas *Accuracy* y Kappa. Esta información se presenta en la Figura 4.4 en donde se puede comparar fácilmente el desempeño de cada ensamble.

Tabla 4.4 Desempeño de los ensambles *C5.0* y
Generalized Boosted Regression Models (GBM) para el conjunto de datos DMRE

Número de remuestreos: 30	
Ensamble	*Accuracy*
C5.0	0.9487663
GBM	0.9442354
	Kappa
C5.0	0.8968537
GBM	0.8876141

Fuente: Elaboración propia.

Figura 4.4 Desempeño de los ensambles C5.0 y *Generalized Boosted Regression Models* (GBM) para el conjunto de datos DMRE



Fuente: Elaboración propia.

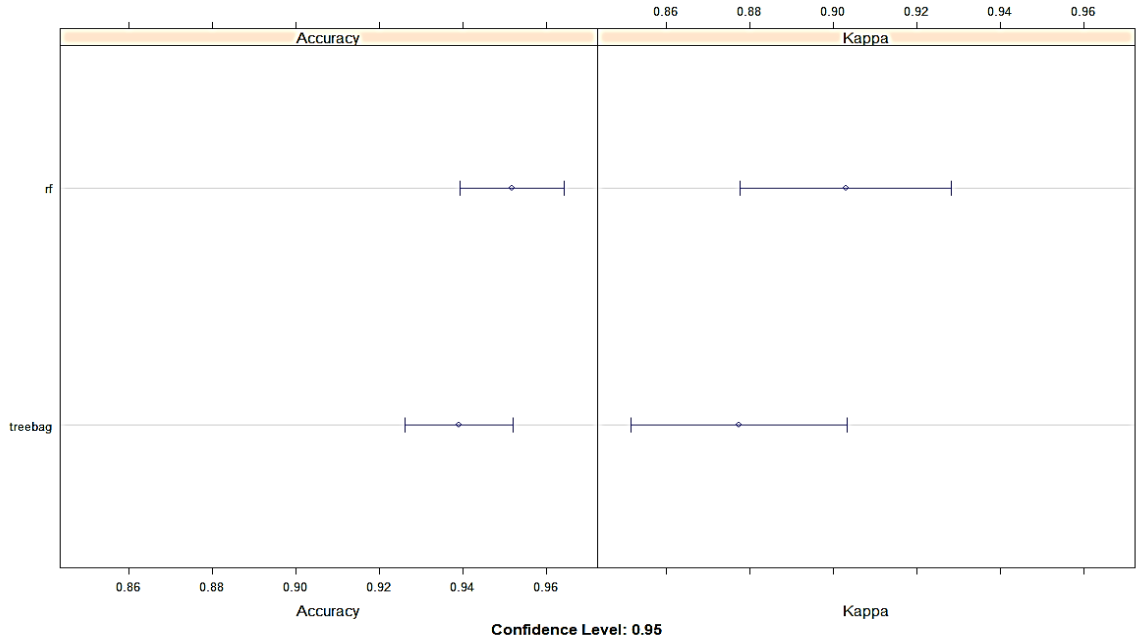
Ensambles Bagging en el conjunto de datos DMRE

La evaluación del desempeño de los ensambles *Árboles de decisión y Random Forest* aplicados al conjunto de datos DMRE, se muestran en la Tabla 4.5. El ensamble generado por el algoritmo *Random Forest* tiene un desempeño ligeramente mejor que *Árboles de decisión* en términos de las métricas Exactitud y Kappa. La Figura 4.5, permite visualizar el rendimiento de ambos ensambles.

Tabla 4.5 Desempeño de los ensambles *Árboles de decisión y Random Forest* para el conjunto de datos DMRE

Número de re-muestreos: 30	
Accuracy	
Árboles de decisión	0.9390510
Random Forest	0.9518303
Kappa	
Árboles de decisión	0.8773611
Random Forest	0.9029772

Figura 4.5 Comparación gráfica del desempeño de los ensambles *Árboles de decisión y Random Forest* para el conjunto de datos DMRE



Fuente: Elaboración propia

Ensamble apilado para el conjunto de datos DMRE

En el apilamiento, cada algoritmo toma las salidas de submodelos como entrada e intenta aprender cómo combinar mejor las predicciones de entrada para hacer una mejor predicción de salida.

En este trabajo se probaron cinco modelos en un ensamble apilado:

i) Análisis discriminante lineal (LDA),
ii) Particionamiento recursivo y árboles de regresión (RPART),
iii) Regresión logística (a través del modelo lineal generalizado o GLM),
iv) Vecinos k-más cercanos (KNN),
v) *Support Vector Machine* con una función de núcleo de base radial (SVM Radial).

Los resultados de la clasificación del conjunto de datos DMRE se muestran en la Tabla 4.6., en donde el mejor evaluado en términos de *Accuracy* y Kappa es SVMRADIAL.

Tabla 4.6 Desempeño de los modelos usados para construir el ensamble apilado para DMRE

Número de re-muestreos: 30	
Ensambles	*Accuracy*
LDA	0.9192483
RPART	0.8816862
GLM	0.9244480
KNN	0.8413914
SVMRADIAL	0.9418969
	Kappa
LDA	0.8361419
RPART	0.7587358
GLM	0.8476054
KNN	0.6819801
SVMRADIAL	0.8831916

Fuente: Elaboración propia.

En un ensamble por apilamiento es deseable que las predicciones hechas por los submodelos tengan baja correlación. Las correlaciones entre los modelos usados se presentan en la Tabla 4.7.

Los modelos más correlacionados son RPART y LDA (0.7037632). Para que una correlación se considere alta, debe ser mayor que 0.75, por lo que los modelos probados en este experimento son válidos para combinar sus predicciones y obtener buenos resultados.

Tabla 4.7 Correlaciones entre los modelos que conforman el ensamble apilado para clasificar el conjunto de datos DMRE

	LDA	RPART	GLM	KNN	SVMRADIAL
LDA	1	0.7037632	0.6721262	0.2945457	0.6954712
RPART	0.7037632	1	0.3790137	0.1817712	0.5478317
GLM	0.6721262	0.3790137	1	0.3971783	0.5800016
KNN	0.2945457	0.1817712	0.3971783	1	0.4984396
SVMRADIAL	0.6954712	0.5478317	0.5800016	0.4984396	1

Fuente: Elaboración propia.

Los resultados obtenidos al combinar los modelos antes evaluados por medio de GLM (*Generalized Linear Model*) son:

Accuracy	Kappa	Sensibilidad	Especificidad
0.9479887	0.8952457	0.9373347	0.9556403

Estos resultados mejoraron ligeramente con respecto a los valores obtenidos por SVMRADIAL, que fue el modelo mejor evaluado de los que integran el ensamble.

4.1.6 **Selección del ensamble adecuado para los conjuntos de datos DMRE**

Con base en los resultados obtenidos en los tres enfoques de ensambles probados se encontró que el que mejor desempeño tiene es el ensamble Bagging con Random Forest, de acuerdo con las métricas *Accuracy* y Kappa.

Tabla 4.8 Comparación del desempeño de los ensambles probados para DMRE

Ensamble	*Accuracy*
C5.0	0.9487663
Random Forest	**0.9518303**
Ensamble apilado con GLM	0.9479887
	Kappa
C5.0	0.8968537
Random Forest	**0.9029772**
Ensamble apilado con GLM	0.8952457

Fuente: Elaboración propia

Ambas métricas resultan consistentes, pues tanto Kappa como *Accuracy* se mueven en la misma dirección para los ensambles ensayados. Esto se debe a que la definición de Kappa está estrechamente relacionada con la definición de *Accuracy*.

Sin embargo, en los ejemplos estudiados se encuentra que Kappa es más adecuado que *Accuracy* porque para un problema de clasificación con conjuntos de datos desequilibrados, *Accuracy* asigna un peso enorme en la clase mayoritaria y un peso muy pequeño en la clase minoritaria. Esto puede llevar a conclusiones erróneas sobre el rendimiento del sistema.

Dado que el coeficiente Kappa considera aciertos casuales, se considera una medida más sólida que *Accuracy*. Kappa es una evaluación que se basa en la diferencia entre el acuerdo real en la matriz de errores y el acuerdo de probabilidad.

Por la razón antes expuesta, se consideró Kappa para decidir que el ensamble adecuado para el estudio del conjunto de datos DMRE es *Random Forest* que, en este caso, coincide con la tendencia de *Accuracy*.

4.1.7 Presentación interpretable de resultados para el conjunto de datos DMRE

La presentación de resultados de manera interpretable, se consideran en tres niveles, de la interpretabilidad en los datos, de los algoritmos utilizados y de las predicciones.

i. La interpretabilidad a nivel de datos implica la presentación de los datos considerados en los experimentos, con el fin de dar claridad a los usuarios finales. Esto se logra definiendo claramente cómo se han seleccionado los datos y las variables tomadas en cuenta. Esta información se presentó en la definición de la población estudiada, en la sección 3.2.
ii. La interpretabilidad para los algoritmos se presentó en la sección 2.8, en donde se presentaron los ensambles que se aplicarían al conjunto de datos en estudio.
iii. La interpretabilidad de las predicciones se presenta a continuación. Para la interpretabilidad global se utilizan árboles de decisión, reglas de inferencia y nomogramas e histogramas. Para lograr la interpretabilidad local, se presentan gráficas elaboradas con base en LIME (*Local Interpretability Model-Agnostic Explanations*).

Con el fin de hacer entendibles los resultados obtenidos se presenta en forma gráfica la forma en que se hacen los razonamientos para llegar a las predicciones propuestas por los

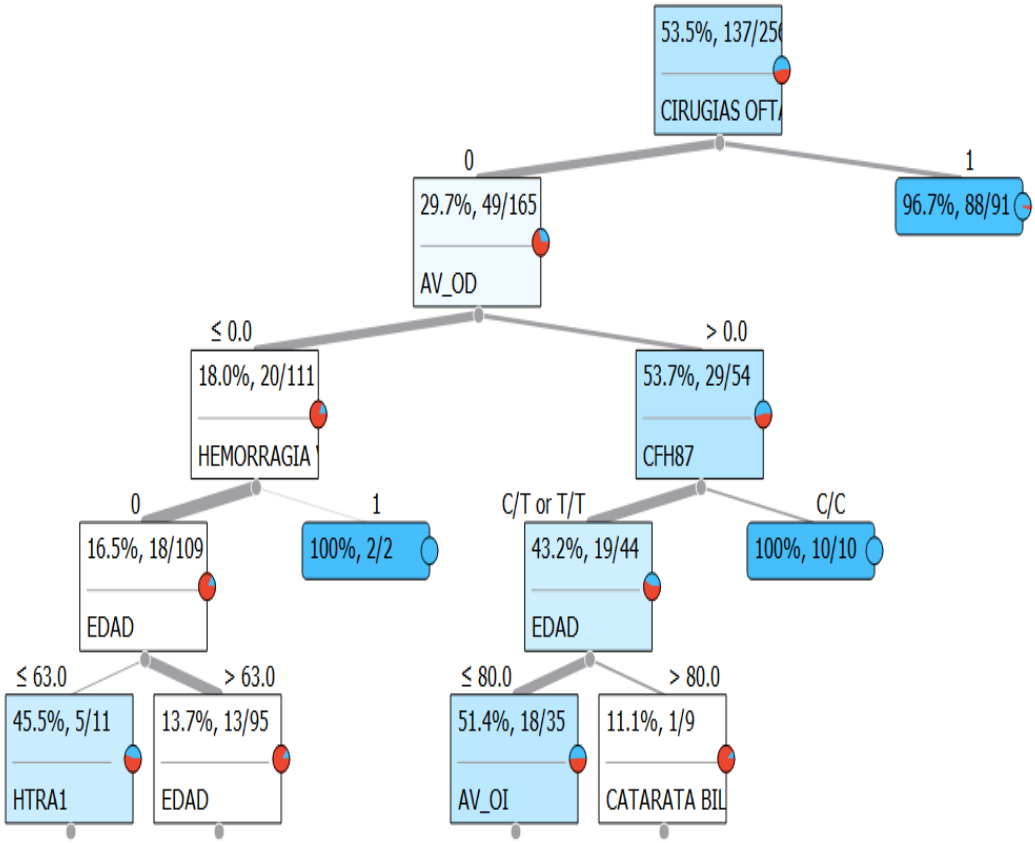
ensambles utilizados. En este escenario de aplicación se hace uso de las que se describen a continuación.

Una representación esquemática que facilita la explicación de las predicciones obtenidas son los árboles de decisión. Estos tienen la ventaja de que los cursos de acción están bien definidos, lo que da claridad al usuario acerca de los pasos seguidos para llegar a las predicciones. El uso de árboles de decisión no implica pérdida considerable en la precisión de los resultados a cambio de claridad en la forma de llegar a los resultados.

Así, el árbol de decisión mostrado en la Figura 4.687. A continuación, se presenta la variable edad, seguida de cataratas bilaterales. En cada caso se muestra la probabilidad de padecer la enfermedad, al recorrerse las ramas del árbol por el lado derecho.

indica que las cirugías oftalmológicas son la variable más relevante para la predicción, seguida de la agudeza visual en el ojo izquierdo, la presencia del polimorfismo CFH87. A continuación, se presenta la variable edad, seguida de cataratas bilaterales. En cada caso se muestra la probabilidad de padecer la enfermedad, al recorrerse las ramas del árbol por el lado derecho.

Figura 4.6 Árbol de decisión para mostrar la forma de llegar a predicciones para el conjunto de datos DMRE



Fuente: Elaboración propia.

Otra forma de explicar la forma de llegar a las predicciones son los nomogramas. Un nomograma, por definición es un instrumento que representa simultáneamente el conjunto de las variables que definen determinado problema y el rango de sus soluciones.

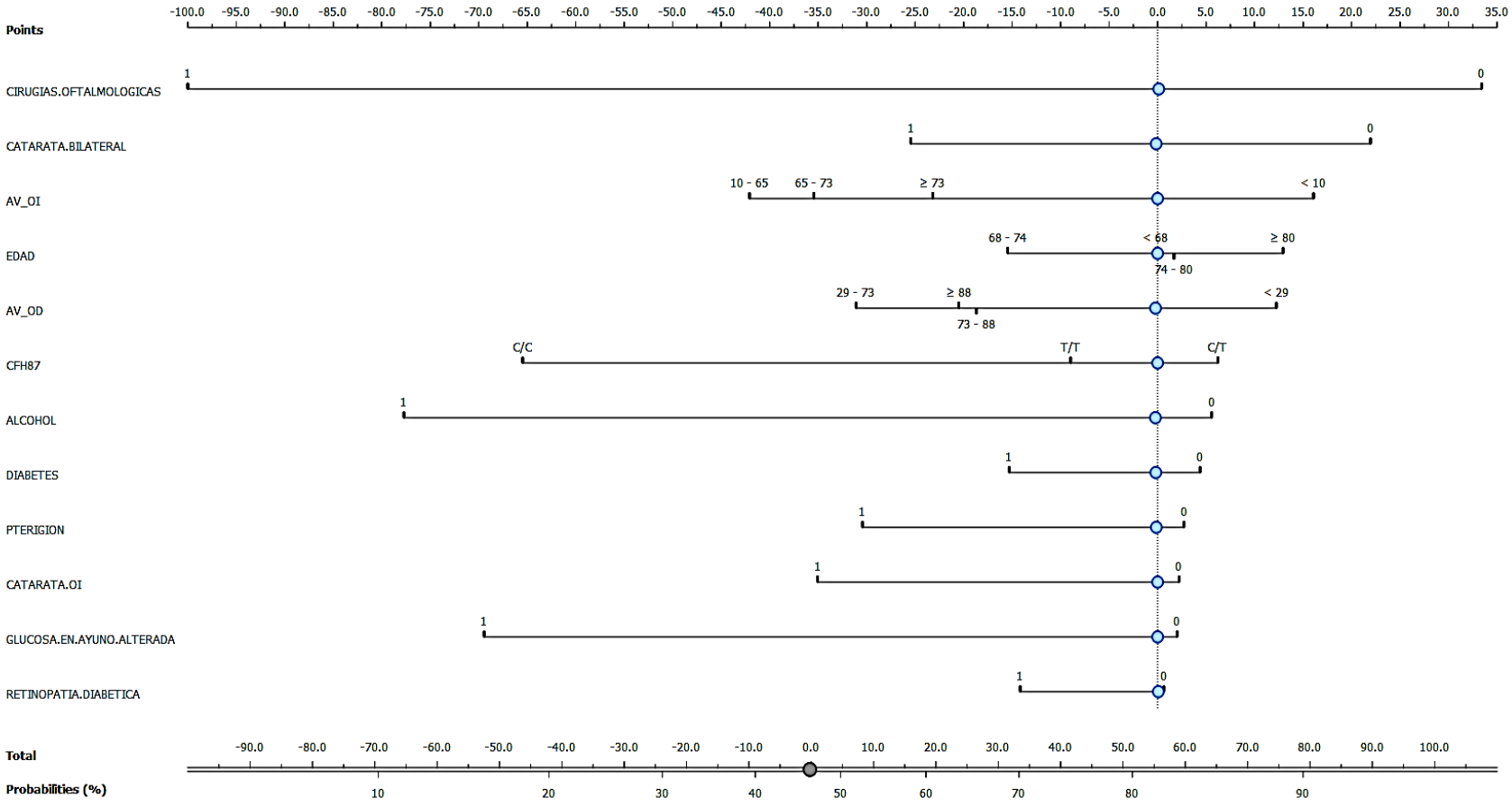
De esta forma, el nomograma mostrado en la Figura 4.7, presenta el listado de las variables más relevantes y permite estimar los valores con que cada variable influye en la probabilidad de que un individuo padezca la enfermedad.

Los valores de cada variable están indicados por medio de un círculo azul y las probabilidades de padecer la enfermedad están indicadas con un círculo gris en la escala inferior.

Este instrumento permite visualizar gráficamente la influencia que pueden tener los cambios en los variables mostradas. En este caso, se muestra que las probabilidades de padecer la enfermedad son del 45% para la combinación de valores indicados en cada variable.

La precisión de un nomograma es limitada, pues está determinada por la forma en que se alinean y perciben los puntos que componen las escalas de valores para cada variable. Sin embrago, son una forma útil y sencilla de mostrar cómo influyen los cambios en las variables en el resultado final.

Figura 4.7 Nomograma para el conjunto de datos DMRE



Fuente: Elaboración propia.

Con el fin de representar la forma en que se llegó a las predicciones, se hace uso de las reglas de inferencia. La estructura de las reglas de inferencia tiene dos partes, la premisa y la conclusión, que se escribe: "Si *premisa*, entonces *conclusión*". En donde las premisas son enunciadas con base en los valores que tienen las variables y la conclusión es la predicción de si el sujeto padece o no la enfermedad, esto es, casos o controles.

Las reglas muestran la forma natural en que se expresan los expertos, es decir, se condiciona la existencia de enfermedad a si se tienen los factores de riesgo encontrados por la clasificación.

En la Figura 4.8 se muestran las reglas de inferencia que pueden guiar a los usuarios del área médica para aclarar la forma en que se llega a los resultados. En la primera columna se muestran las premisas, es decir, las variables; en la segunda la conclusión, en la tercera columna está descrita la proporción de casos: controles en el conjunto de datos; en la última columna se describe la probabilidad de pertenencia al grupo de casos o controles.

Sólo se muestran las reglas con una, dos y tres premisas, para favorecer la claridad de las inferencias. En el primer renglón la premisa "TRUE", indica que la clase CASO es verdadera con una proporción de 119, sobre 137 elementos de la clase CONTROL (47% a 53%).

En los renglones subsecuentes, en la columna de las premisas se enlistan las variables con los valores que corresponden a la predicción (conclusión). El listado comienza con reglas de una premisa y termina con reglas que incluyen tres premisas, para apoyar la comprensibilidad.

Figura 4.8 Reglas de inferencia generadas para el conjunto de datos DMRE

		THEN class	Distribution	Probabilities [%]
TRUE	→	GRUPO=CONTROL	[119, 137]	47 : 53
TABQ≠N	→	GRUPO=CASO	[52, 0]	98 : 2
DESPRENDIMIENTO DE VITREO POSTERIOR COMPLETO OD≠N	→	GRUPO=CASO	[1.0, 0.0]	67 : 33
ECTOPRION≠N	→	GRUPO=CASO	[1.4, 0.0]	71 : 29
FRACTURA COROIDEA≠N	→	GRUPO=CONTROL	[0, 1]	33 : 67
BIEFARITIS≠N	→	GRUPO=CONTROL	[0.0, 1.0]	33 : 67
AV_OI≥100.0	→	GRUPO=CONTROL	[0.0, 3.0]	20 : 80
HTRAT=C/C	→	GRUPO=CONTROL	[11.0, 1.1]	85 : 15
DISLIPIDEMIA≠N	→	GRUPO=CONTROL	[9.0, 1.2]	82 : 18
EDAD≥87.0	→	GRUPO=CONTROL	[16.0, 1.3]	88 : 12
TABQ≠N AND CFH28≠C/C	→	GRUPO=CASO	[18.7, 0.0]	95 : 5
TABQ≠N AND CFH28=C/T	→	GRUPO=CASO	[9.7, 0.0]	91 : 9
TABQ≠N AND CFH28≠C/T	→	GRUPO=CASO	[20.7, 0.0]	96 : 4
TABQ≠N AND CFH28=T/T	→	GRUPO=CASO	[2.4, 0.0]	77 : 23
TABQ≠N AND CFH28≠T/T	→	GRUPO=CASO	[18.9, 0.0]	95 : 5
TABQ≠N AND CFH87≠C/C	→	GRUPO=CASO	[14.5, 0.0]	94 : 6
TABQ≠N AND CFH87=C/T	→	GRUPO=CASO	[9.6, 0.0]	91 : 9
TABQ≠N AND CFH87≠T/T	→	GRUPO=CASO	[7.1, 0.0]	89 : 11
TABQ≠N AND HTRAT≠C/C	→	GRUPO=CASO	[5.4, 0.0]	86 : 14
TABQ≠N AND HTRAT=C/T	→	GRUPO=CASO	[3.0, 0.0]	80 : 20
TABQ≠N AND HTRAT≠T/T	→	GRUPO=CASO	[2.2, 0.0]	76 : 24
TABQ≠N AND HTA=N	→	GRUPO=CASO	[1.1, 0.0]	68 : 32
TABQ≠N AND HTA≠N	→	GRUPO=CASO	[1.3, 0.0]	69 : 31
TABQ≠N AND ALCOHOL=N	→	GRUPO=CASO	[1.3, 0.0]	69 : 31
TABQ≠N AND GLUCOSA EN AYUNO ALTERADA=N	→	GRUPO=CASO	[1.2, 0.0]	69 : 31
ALCOHOL≠N AND AGUJERO EN MACULA≠N	→	GRUPO=CASO	[1.0, 0.0]	67 : 33
CATARATA OD≠N AND HTRAT=C/C	→	GRUPO=CASO	[1.0, 0.0]	67 : 33
HTA=N AND TABQ=N	→	GRUPO=CASO	[1.2, 100.0]	2 : 98
CIRUGIAS OFTALMOLOGICAS≠N AND EDAD≤81.0	→	GRUPO=CASO	[1.0, 53.0]	4 : 96
CIRUGIAS OFTALMOLOGICAS≠N AND TABQ=N	→	GRUPO=CASO	[1.3, 60.0]	4 : 96
XEROSIS AO≠N AND EDAD≥96.0	→	GRUPO=CASO	[1.0, 0.0]	67 : 33
CFH87≠C/T AND EDAD≥89.0	→	GRUPO=CASO	[1.4, 0.0]	70 : 30
AV_OI≥20.0 AND HTRAT=C/C	→	GRUPO=CASO	[1.0, 0.0]	67 : 33
EDAD≤77.0 AND EDAD≥77.0	→	GRUPO=CASO	[2.0, 0.0]	75 : 25
AV_OI≥100.0 AND CFH87=C/T	→	GRUPO=CONTROL	[0.0, 1.4]	29 : 71
AV_OI≥100.0 AND HTRAT=T/T	→	GRUPO=CONTROL	[0.0, 1.2]	31 : 69
HTA=N AND HTRAT=C/C	→	GRUPO=CONTROL	[0.0, 5.0]	14 : 86
HTA=N AND CATARATA OD≠N	→	GRUPO=CONTROL	[0.0, 3.0]	20 : 80
HTA=N AND AGUJERO EN MACULA≠N	→	GRUPO=CONTROL	[0.0, 1.0]	33 : 67
RETINOPATIA DIABETICA≠N AND CFH28=T/T	→	GRUPO=CONTROL	[0.0, 1.0]	33 : 67
RETINOPATIA DIABETICA≠N AND CFH87=C/C	→	GRUPO=CONTROL	[0.0, 1.0]	33 : 67
HTRAT=C/C AND CIRUGIAS OFTALMOLOGICAS≠N	→	GRUPO=CONTROL	[0.0, 1.5]	29 : 71

Fuente: Elaboración propia

Si bien, es indispensable explicar de manera general la forma en que se llega a las predicciones obtenidas para el conjunto de datos DMRE, es muy importante dar explicaciones para los individuos seleccionados. Esto es por tratarse de datos provenientes de personas.

En consecuencia, se presenta una representación gráfica, elaborada con base en el modelo LIME (*Local Interpretable Model-agnostic Explanations*). En la Figura 4.9 se presenta una imagen en la que se muestra la forma de llegar a la conclusión de si un sujeto pertenece al grupo 0 (control) o al grupo 1 (caso). Se muestran los factores de riesgo en color azul, del lado derecho, y los factores protectores en rojo, del lado izquierdo.

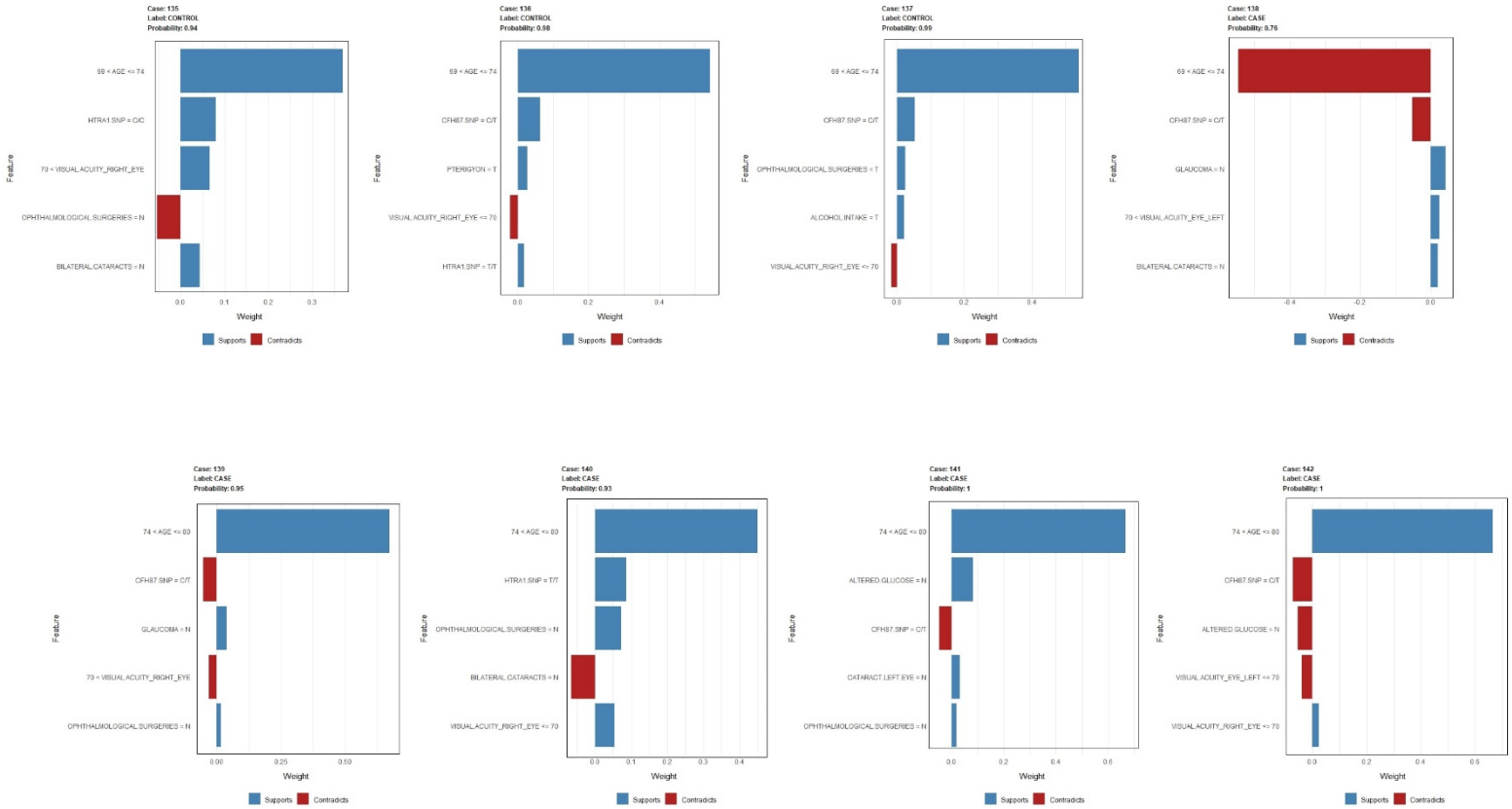
Esta representación gráfica de los resultados permite que estos se verifiquen en instancias particulares con las variables que resulten de mayor importancia para el usuario de los resultados. El gráfico se interpreta como se describe a continuación.

Para el primer elemento de la primera columna de la imagen, se muestra que para el individuo identificado con el número 135, cuya clase es “0” (control), el factor de riesgo identificado es el polimorfismo CFH87 con la combinación de bases C/C, mientras que los factores protectores son la presencia de cirugías oftálmicas (se toma un rango entre 0 y 1, por lo que >0.75, indica que el sujeto tiene registradas cirugías oftálmicas) y edad menor o igual a 68 años.

Para el segundo caso de la primera columna, el individuo identificado como caso 138, su clase es “1” (caso), los factores de riesgo identificados son: el polimorfismo CFH87 con combinación de bases C/T y el polimorfismo HTRA con combinación de bases C/T. EL factor protector es la ausencia de cirugías oftálmicas.

Como se puede apreciar, la interpretabilidad local no es necesariamente fiel a la global y viceversa. Esto no implica que haya contradicción entre ellas.

Figura 4.9 Resultados para instancias seleccionadas (Interpretabilidad Local)



Fuente: Elaboración propia

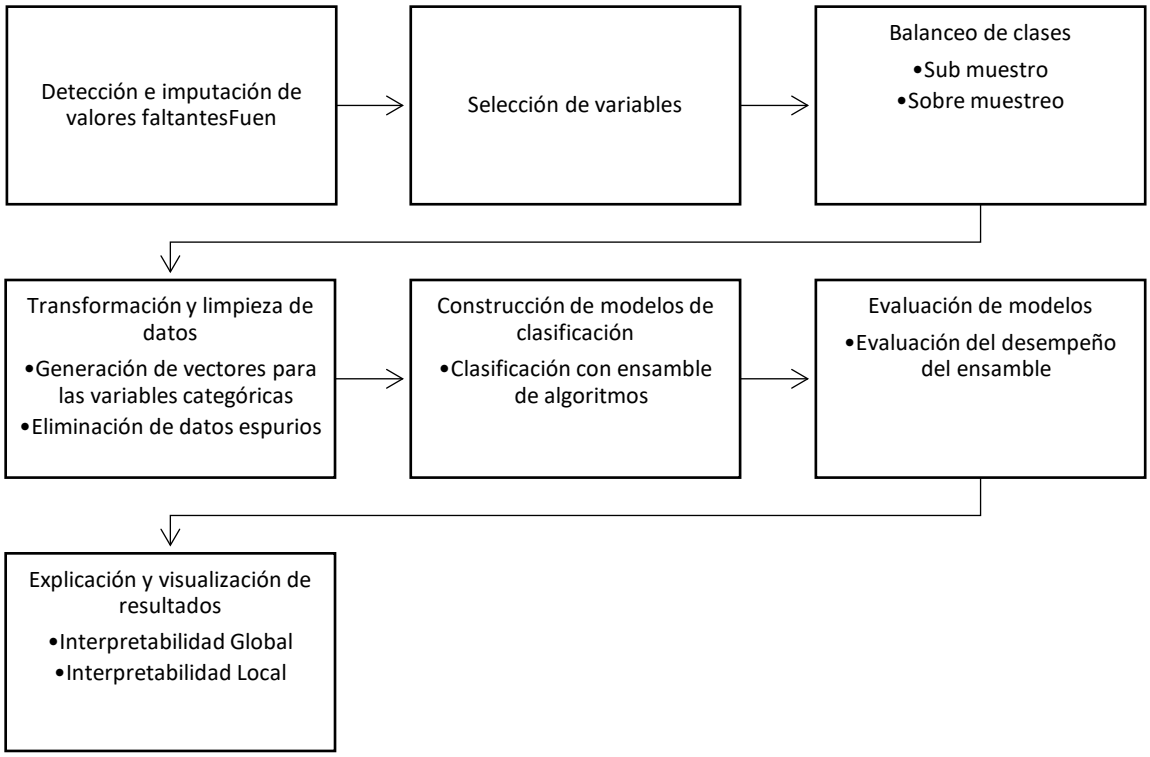
De esta manera, las explicaciones globales buscan que, a partir de la generalidad se valide la generación de conocimiento; y en las explicaciones locales se apoye en el tratamiento de una persona en particular.

A continuación, se describe la limpieza y transformación de los datos, el uso de métodos de aprendizaje automático para encontrar las relaciones entre las variables médicas y socioeconómicas y la Preeclampsia en mujeres que viven en Nueva Jersey. De la misma manera, se muestran los resultados por medio de técnicas de interpretabilidad para explicar a los expertos en salud cómo se llegó a los resultados.

4.2 Segundo escenario de aplicación: Preeclampsia

El segundo escenario de aplicación es Preeclampsia. En este caso, se comenzó a trabajar a partir de la segunda fase de la metodología descrita en el capítulo 2 (ver Figura 4.10). El conjunto de datos resulta de interés por su alto grado de desbalanceo y por las implicaciones que esta enfermedad tiene en la población a nivel global, así como la necesidad de presentar los resultados de manera interpretable.

Figura 4.10 Metodología para determinar los factores de riesgo asociados a Preeclampsia mediante el enfoque de aprendizaje automático.



Fuente: Elaboración propia

4.2.1 **Descripción de la base de datos**

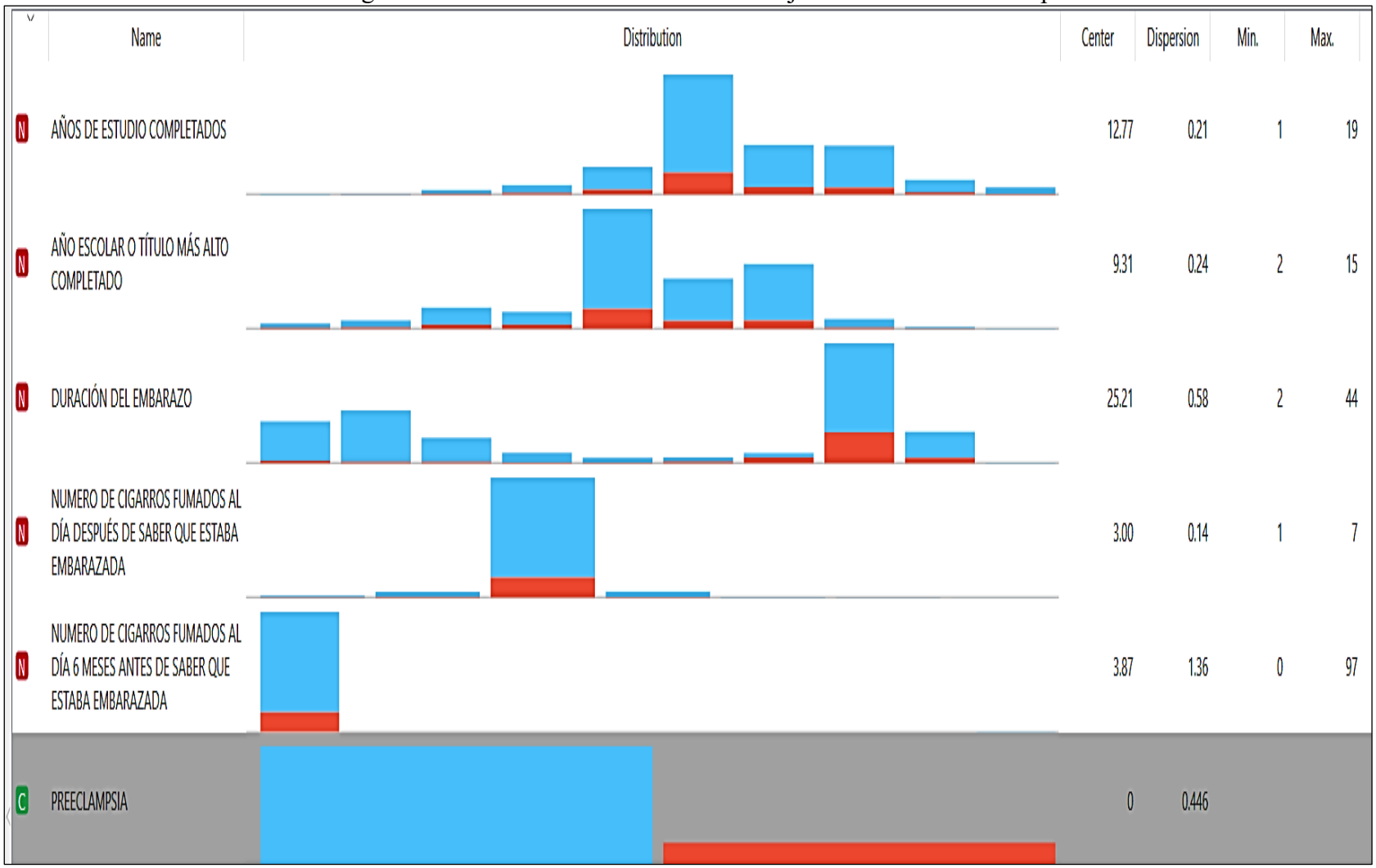
Los datos se obtuvieron de la iniciativa “Children's Futures” (Inter-university Consortium for Political and Social Research, 2008), que fue diseñada para mejorar la salud y el bienestar de los niños desde el nacimiento hasta los tres años en Trenton, Nueva Jersey a través de tres estrategias principales: (1) Mejorar el acceso a la atención prenatal y fortalecer la crianza efectiva; (2) Mejorar la calidad del cuidado infantil; y (3) Fortalecer y mantener una participación positiva de los padres en la vida de sus hijos.

Como parte de la iniciativa, se recopilaron datos simultáneamente para evaluar la efectividad de esta. La recopilación de datos incluyó una encuesta de referencia de la comunidad de Trenton realizada en 2002 y encuestas de proveedores de cuidado infantil de Trenton realizadas en 2003, 2004 y 2005. Además, se obtuvieron registros de nacimientos de Trenton, Camden y Newark del estado de Nueva Jersey Departamento de salud. Los datos de las encuestas y registros de nacimientos se publicaron como ICPSR 21640: Evaluación del futuro de los niños: Mejora de los resultados de salud y desarrollo para niños en Trenton, Nueva Jersey, 2001-2005.

El conjunto de datos incluye 1640 registros, con 12.4% de datos faltantes, que se sustituyeron por la mediana o el promedio para las variables categóricas y las numéricas, respectivamente.

Las variables numéricas del conjunto de datos se presentan en la Figura 4.11 , en donde se señalan en color azul los casos de personas que no padecen la enfermedad y en rojo las que sí sufren de ella.

Figura 4.11 Variables numéricas en el conjunto de datos Preeclampsia



Fuente: Elaboración propia.

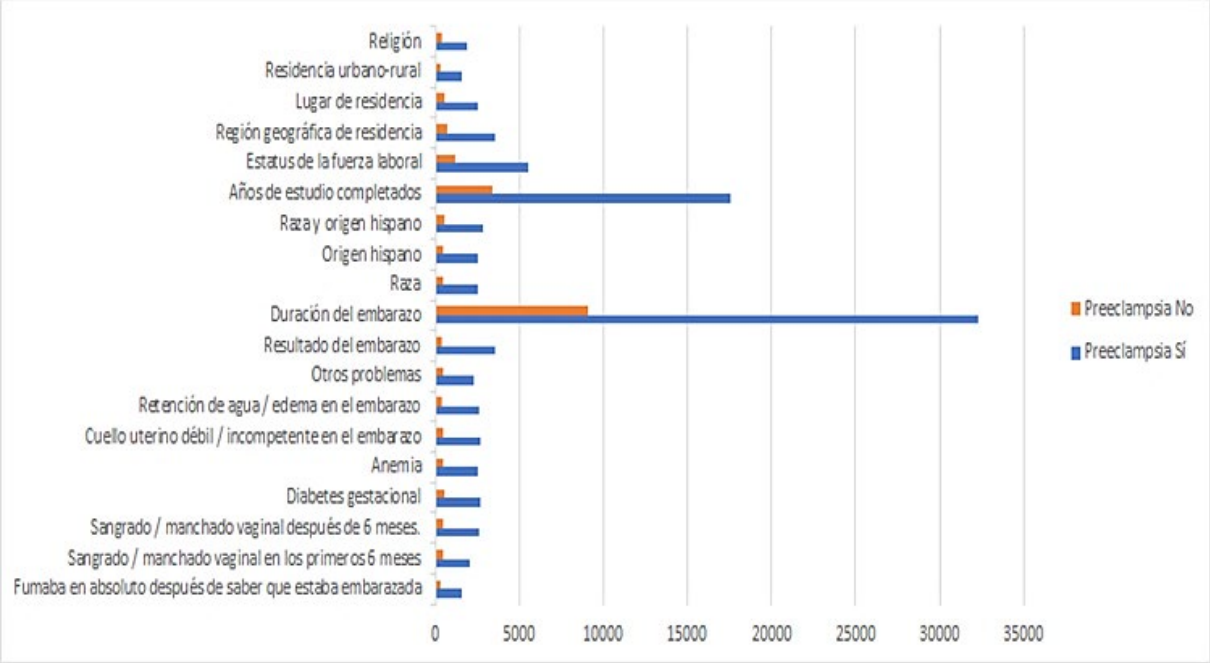
Las variables categóricas del conjunto de datos se muestran en la Tabla 4.9 y Figura 4.12, en donde se hace un recuento de las que están presentes en las instancias de personas que padecen Preeclampsia y las que no la padecen.

Tabla 4.9 Variables categóricas para el conjunto de datos Preeclampsia

Variables	Preeclampsia	
	Sí	No
Preeclampsia	1371	269
Fumaba en absoluto después de saber que estaba embarazada	1617	323
Sangrado / manchado vaginal en los primeros 6 meses	2044	445
Sangrado / manchado vaginal después de 6 meses.	2635	508
Diabetes gestacional	2657	511
Anemia	2546	476
Cuello uterino débil / incompetente en el embarazo	2692	505
Retención de agua / edema en el embarazo	2610	414
Otros problemas	2312	466
Resultado del embarazo	3575	379
Duración del embarazo	32271	9070
Raza	2534	459
Origen hispano	2520	491
Raza y origen hispano	2870	578
Años de estudio completados	17581	3368
Estatus de la fuerza laboral	5506	1211
Región geográfica de residencia	3569	706
Lugar de residencia	2520	510
Residencia urbano-rural	1539	310
Religión	1912	361

Fuente: Elaboración propia.

Figura 4.12 Variables categóricas para el conjunto de datos Preeclampsia, agrupadas por personas que padecen la enfermedad y quienes no la padecen



Fuente: Elaboración propia.

4.2.2 **Selección de las variables más relevantes para Preeclampsia**

Como primera acción se procedió a revisar si el conjunto de datos tenía valores faltantes. La cantidad de faltantes fue de 12.4% que se sustituyeron por la media o la mediana, si la variable era numérica o categórica, respectivamente.

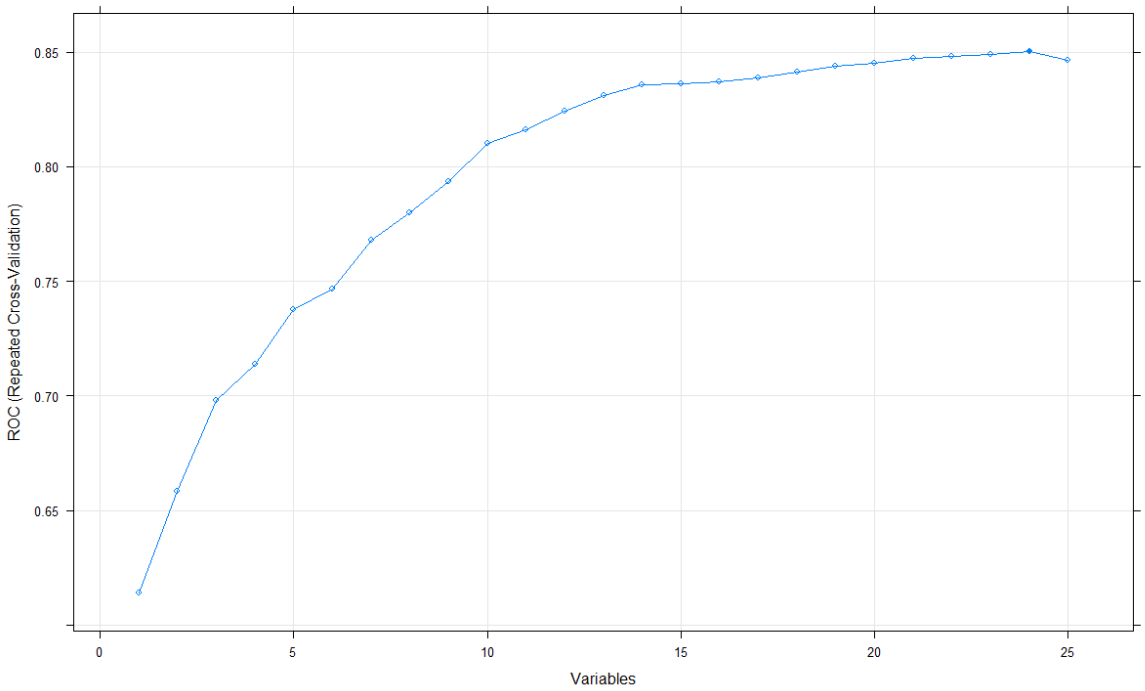
Una vez que la base de datos fue pre procesada se realizó la selección de variables por medio RFE con *Random Forest.* En donde se clasificó de acuerdo con su importancia para el modelo. La precisión de los modelos generados es la métrica de evaluación para determinar el poder predictivo del conjunto de variables en el proceso de clasificación.

El algoritmo RFE probó todas las combinaciones posibles y las almacenó en una lista de combinación de variables y su rendimiento. Para cada iteración, todas las variables se clasifican nuevamente.

Para seleccionar las variables más relevantes, a pesar del problema del desequilibrio entre clases, se utilizó la curva de características operativas del receptor (ROC) para el método de validación cruzada *Leave One Out* como se muestra en la Figura 4.13 (la curva

ROC penaliza dicha condición al limitar el valor máximo posible a lo largo del eje). Con base en la curva ROC se decidió usar las 25 variables para continuar con los pasos subsecuentes para la clasificación del conjunto de datos.

Figura 4.13 Conjuntos de variables analizadas ordenadas por medio de ROC para Preeclampsia



Fuente: Elaboración propia.

Las variables que conforman el conjunto de datos Preeclampsia están enlistadas en la Tabla 4.10 , de acuerdo con la Disminución media del índice de Gini (MDGI). El valor MDGI de cada variable se expresa en el rango [0, 100].

Las variables más importantes son la duración del embarazo completado en semanas (PRGLNGTH), pobreza (POVERTY), Estatus de fuerza laboral, la retención de agua / edema en el embarazo (SWLNANKL), la educación (años de escolaridad completados) (EDUCAT), estado de la fuerza laboral (LABORFOR).la escuela o el grado escolar más alto (HIEDUC), Número de cigarros fumados al día 6 meses antes de saber que estaba embarazada (PRIORSMK).

Tabla 4.10. Evaluación de los atributos del conjunto de datos Preeclampsia de acuerdo con el de Gini (MDGI)

Variables		MDGI
PRGLNGTH	Duración del embarazo	27.799686
POVERTY	Nivel de pobreza de ingresos	25.802818
SWLNANKL	Retención de agua / edema en el embarazo	19.419097
EDUCAT	Años de estudio completados	11.144021
LABORFOR	Estatus de la fuerza laboral	10.898961
HIEDUC	Año escolar o título más alto completado	10.355996
PRIORSMK	Número de cigarros fumados al día 6 meses antes de saber que estaba embarazada	9.398163
OUTCOME	Resultado del embarazo	8.427174
REGION	Región geográfica de residencia	8.115532
RELIGION	Religión	7.606145
NEWPR	Nuevo embarazo	6.800513
METRO	Lugar de residencia	6.679053
OTHRPROB	Otros problemas	6.291588
RACE	Raza	4.490506
VGBLDFST	Sangrado / manchado vaginal en los primeros 6 meses	4.171894
HISPRACE	Raza y origen hispano	4.099988
ANEMIA	Anemia	3.856929
GESTDBTS	Diabetes gestacional	3.712030
WEAKCRVX	Cuello uterino débil / incompetente en el embarazo	3.109812
NPOSTSMK	Numero de cigarros fumados al día después de saber que estaba embarazada	2.964841
POSTSMKS	Fumaba en absoluto después de saber que estaba embarazada	2.891663
VGBLDLST	Sangrado / manchado vaginal después de 6 meses	2.799006

Fuente: Elaboración propia

4.2.3 **Balanceo del conjunto de datos Preeclampsia**

El conjunto de datos está desbalanceado, con 269 instancias para la clase positiva y 1371 para la clase negativa y 26 variables. La tasa de desbalanceo es de 0.1962071, la proporción entre los casos y controles es de 0.1640244 para las personas enfermas, y 0.8359756 para quienes no padecen la enfermedad.

El conjunto de datos Preeclampsia tiene una tasa de desbalanceo de 0.1962071, que se resuelve por medio del algoritmo SMOTE. Se hicieron cuatro experimentos cambiando las proporciones de sobre muestreo/ sub muestro (ver

Tabla 4.11):

i) la proporción 4/1 aumentó la cantidad de instancias de la clase minoritaria a 1345 y disminuyó a 1076 las instancias la clase mayoritaria, con tasa de desbalanceo (IR) de 0.8,

ii) la proporción 2/2 aumentó la cantidad de instancias de la clase minoritaria a 807 y disminuyó a 1076 las instancias de clase mayoritaria con tasa de desbalanceo (IR) de 0.75,

iii) la proporción 4/2 aumentó la cantidad de instancias de la clase minoritaria a 2152 y disminuyó a 1345 las instancias de clase mayoritaria con tasa de desbalanceo (IR) de 0.625,

iv) la proporción 2/1 aumentó la cantidad de instancias de la clase minoritaria a 538 y disminuyó a 538 las instancias de clase mayoritaria con tasa de desbalanceo (IR) de 1.0.

La tasa de desbalanceo no necesariamente indica la mejor proporción entre las clases. Para decidir acerca de la tasa de muestreo adecuada, se deberá tomar en cuenta la tasa de desbalanceo, así como la pérdida de información que se puede sufrir por submuestreo de la clase mayoritaria.

Tabla 4.11 Instancias sobre y sub muestreadas por medio de SMOTE para el conjunto de datos Preeclampsia

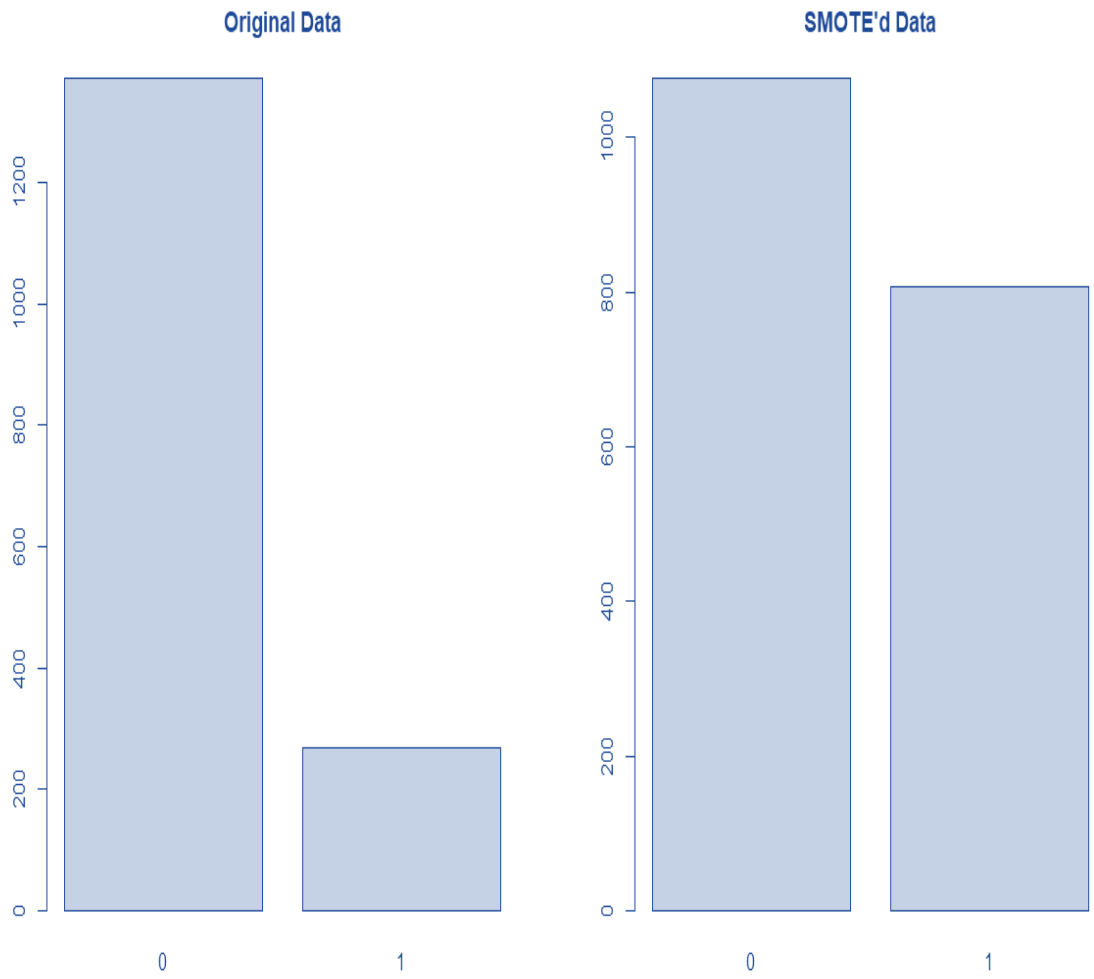
Proporción de sobre muestreo y submuestreo	Casos	Controles	Tasa de desequilibrio
ORIGINAL	269	1371	0.1962071
4/1	1345	1076	0.8
2/2	**807**	**1076**	**0.75**
4/2	2152	1345	0.625
2/1	538	538	1

Fuente: Elaboración propia.

Con el fin de encontrar un conjunto de datos que satisfaga ambas condiciones se decidió tomar la proporción 2/2 con una tasa de desequilibrio de 0.75 que es mucho mejor

que la tasa inicial de 0.1962071. En la Figura 4.14 se muestra gráficamente la proporción de las clases antes y después del sobre y submuestreo.

Figura 4.14 Conjunto de datos Preeclampsia antes y después de aplicar SMOTE con proporción de sobre muestreo 2/2



Fuente: Elaboración propia.

Una vez solventado el problema del desbalanceo de las clases, se procede a la eliminación de los datos ruidosos, como se describe en la siguiente sección.

4.2.4 Eliminación de instancias espurias para el conjunto de datos Preeclampsia

La mayor parte de las variables del conjunto de datos Preeclampsia son categóricas, esto es, 21 variables categóricas y sólo 5 de ellas son numéricas. Con el fin de evitar que las variables categóricas interfieran con la clasificación, pues el algoritmo puede asumir que las enumeraciones, se transformaron los datos para que cada categoría de la variable categórica en cuestión tenga un valor numérico en un vector binario. Esto se realizó por medio del algoritmo de codificación *"One hot"*.

Una vez generadas los vectores correspondientes, se utilizó el algoritmo *Tomek links* para eliminar las instancias de la clase mayoritaria que están muy cerca de instancias de la clase minoritaria. Para el conjunto de datos Preeclampsia, se eliminaron 184 instancias, el 17.1 % de la clase mayoritaria.

Así, ya que se ha resuelto los problemas de desbalanceo de clases, el de las variables categóricas y el de los datos espurios, se prosigue con la clasificación de los conjuntos de datos.

4.2.5 Ensambles de datos para clasificar el conjunto de datos Preeclampsia

El conjunto de datos Preeclampsia tenía un desbalance entre clases mucho mayor que el de DMRE. Los datos se balancearon de forma que la tasa de desbalanceo final es de 0.75 para continuar con la clasificación. Para hacerlo, se usaron tres enfoques distintos de ensambles: *Boosting, Bagging y* Apilamiento (*Stack)* y se comparó el desempeño de cada caso.

Ensambles Boosting en el conjunto de datos Preeclampsia

Los resultados de la clasificación por medio de ensambles *Boosting*, se presentan en la

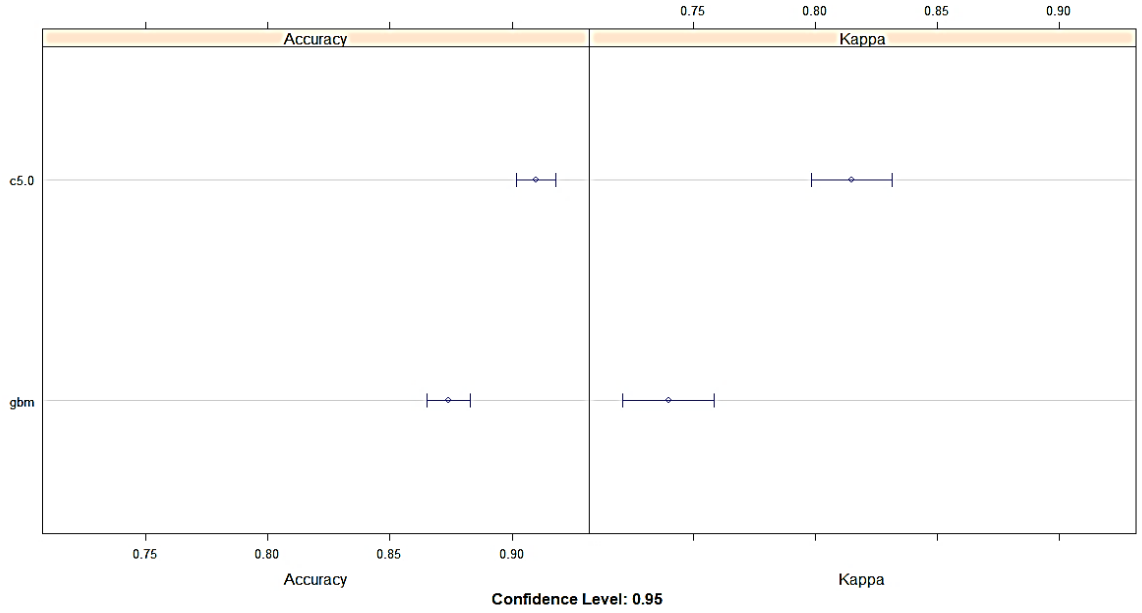
Tabla 4.12 , en donde se evalúa el desempeño de los ensambles C5.0 y GBM. El ensamble C5.0 obtiene un desempeño mejor que GBM, de acuerdo con las métricas *Accuracy* y Kappa. Esta información se presenta en la Figura 4.15, en donde se puede comparar fácilmente el desempeño de cada ensamble.

Tabla 4.12 Desempeño de los ensambles *C5.0* y *GBM* para el conjunto de datos /Preeclampsia

Número de remuestreos: 30	
Accuracy	
C5.0	0.9099017
GBM	0.8741310
Kappa	
C5.0	0.8149243
GBM	0.7399147

Fuente: Elaboración propia

Figura 4.15 Comparación gráfica de los ensambles c5.0 y GBM para el conjunto de datos Preeclampsia



Fuente: Elaboración propia.

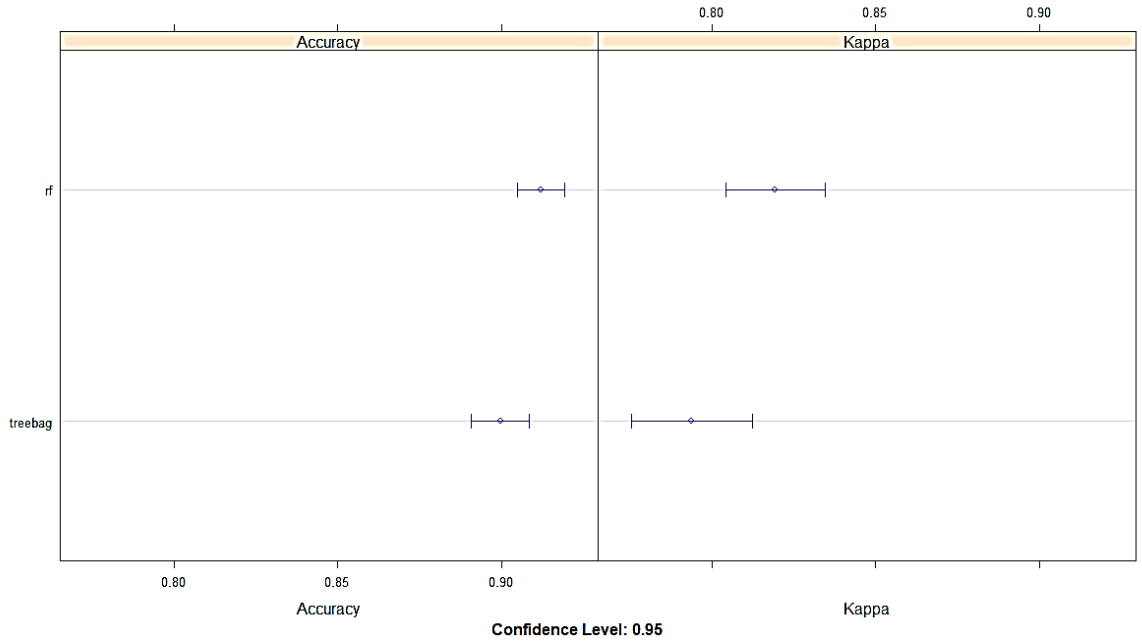
Ensambles Bagging en el conjunto de datos Preeclampsia

La evaluación del desempeño de los ensambles *Árboles de decisión y Random Forest* aplicados al conjunto de datos Preeclampsia, se muestra en la Tabla 4.13. El ensamble generado por el algoritmo *Random Forest* tiene un desempeño mejor que *Árboles de decisión* en términos de las métricas *Accuracy* y Kappa. La Figura 4.16, permite visualizar gráficamente el rendimiento de ambos ensambles.

Tabla 4.13 Desempeño de los ensambles *Árboles de decisión y Random Forest* para el conjunto de datos Preeclampsia

Número de remuestreos: 30	
Accuracy	
Árboles de decisión	0.8998056
Random Forest	0.9121973
Kappa	
Árboles de decisión	0.7937340
Random Forest	0.8193311

Figura 4.16 Desempeño de los ensambles *Árboles de decisión y Random Forest* para el conjunto de datos Preeclampsia



Ensamble apilado para el conjunto de datos Preeclampsia

En el apilamiento, cada algoritmo toma las salidas de submodelos como entrada para aprender las mejores combinaciones para lograr la mejor predicción de salida.

En este trabajo se probaron cinco modelos en un ensamble apilado:

i) Análisis discriminante lineal (LDA),

ii) Particionamiento recursivo y árboles de regresión (RPART),

iii) Regresión logística (a través del modelo lineal generalizado o GLM),

iv) Vecinos k-más cercanos (KNN),

v) *Support Vector Machine* con una función de núcleo de base radial (SVM Radial).

En un ensamble por apilamiento es deseable que las predicciones hechas por los submodelos tengan baja correlación. Las correlaciones entre los modelos usados se presentan en la Tabla 4.14.

Los modelos más correlacionados son GLM y LDA (0.9473323). Para que una correlación se considere alta, debe ser mayor que 0.75, por lo que uno de los modelos GLM ó LDA debe ser removido del ensamble.

Tabla 4.14 Correlaciones entre los modelos que conforman el ensamble apilado para clasificar el conjunto de datos Preeclampsia

	LDA	**RPART**	**GLM**	**KNN**	**SVMRADIAL**
LDA	1	0.6023016	0.9473323	0.04850187	0.7749167
RPART	0.60230162	1	0.5906132	0.15429452	0.5448245
GLM	0.94733229	0.5906132	1	0.05734260	0.7920015
KNN	0.04850187	0.1542945	0.0573426	1	0.1165927
SVM RADIAL	0.77491670	0.5448245	0.7920015	0.11659274	1

Fuente: Elaboración propia.

Los resultados de la clasificación del conjunto de datos Preeclampsia se muestran en la Tabla 4.15, en donde el mejor evaluado en términos de *Accuracy* y Kappa es SVMRADIAL.

Tabla 4.15 Desempeño de los modelos usados para construir el ensamble apilado para el conjunto de datos Preeclampsia

Número de remuestreos: 30	
Accuracy	
LDA	0.7500606
RPART	0.7797768
GLM	0.7436841
KNN	0.7302011
SVMRADIAL	0.8189203
Kappa	
LDA	0.4834341

RPART	0.5465129
GLM	0.4697603
KNN	0.4424217
SVMRADIAL	0.6299330

Fuente: Elaboración propia.

Para evitar el problema generado por los algoritmos altamente correlacionados, se eliminó LDA y se hizo el apilamiento con *Random Forest*.

Accuracy	Kappa	Sensibilidad	Especificidad
0.8514762	0.6940580	0.7901524	0.8948830

Estos resultados mejoraron los valores obtenidos para SVMRADIAL, que fue el modelo mejor evaluado de los que integran el ensamble.

4.2.6 Selección del ensamble adecuado para los conjuntos de datos Preeclampsia

Con base en los resultados obtenidos en los tres enfoques de ensambles probados se encontró que el que mejor desempeño tiene es el ensamble Bagging con Random Forest, de acuerdo con las métricas *Accuracy* y Kappa.

Tabla 4.16 Comparación del desempeño de los ensambles probados para el conjunto de datos Preeclampsia

Accuracy	
C5.0	0.9099017
Random Forest	**0.9121973**
Ensamble apilado con Random Forest	0.8776750
Kappa	
C5.0	0.8149243
Random Forest	**0.8193311**
Ensamble apilado con Random Forest	0.7501822

Fuente: Elaboración propia.

Ambas métricas resultan consistentes, pues tanto Kappa como *Accuracy* se mueven en la misma dirección para los ensambles ensayados. Esto se debe a que la definición de Kappa está estrechamente relacionada con la definición de *Accuracy*.

En el caso de Preeclampsia se encuentra que Kappa es más adecuado que *Accuracy* porque para un problema de clasificación con conjuntos de datos desequilibrados, *Accuracy*

asigna un peso enorme en la clase mayoritaria y un peso muy pequeño en la clase minoritaria. Esto puede llevar a conclusiones erróneas sobre el rendimiento del sistema.

Kappa es una evaluación que se basa en la diferencia entre el acuerdo real en la matriz de errores y el acuerdo de probabilidad. En consecuencia, se consideró Kappa para decidir que el ensamble adecuado para el estudio del conjunto de datos Preeclampsia es *Random Forest* que, en este caso coincide con la tendencia de *Accuracy*.

En ambos escenarios de aplicación *Random Forest (Bagging)* resultó ser el mejor ensamble de clasificadores, a pesar de que el balance entre clases es distinto para cada conjunto de datos.

4.2.7 Presentación interpretable de resultados para el conjunto de datos Preeclampsia

A continuación, se muestran los tres niveles de interpretabilidad para el conjunto de datos Preeclampsia.

La interpretabilidad a nivel de datos se presentó en la sección 3.7.2, en donde cómo se obtuvieron los datos, la importancia de estos y el tipo de variables presentes en el conjunto de datos. La representación gráfica es de gran ayuda para hacer accesibles las explicaciones.

La interpretabilidad a nivel de algoritmos se presentó en la sección 3.5, en donde se explica la forma en que se construyeron los ensambles de clasificadores, los algoritmos de clasificación incluidos en ellos y la forma en que trabajan.

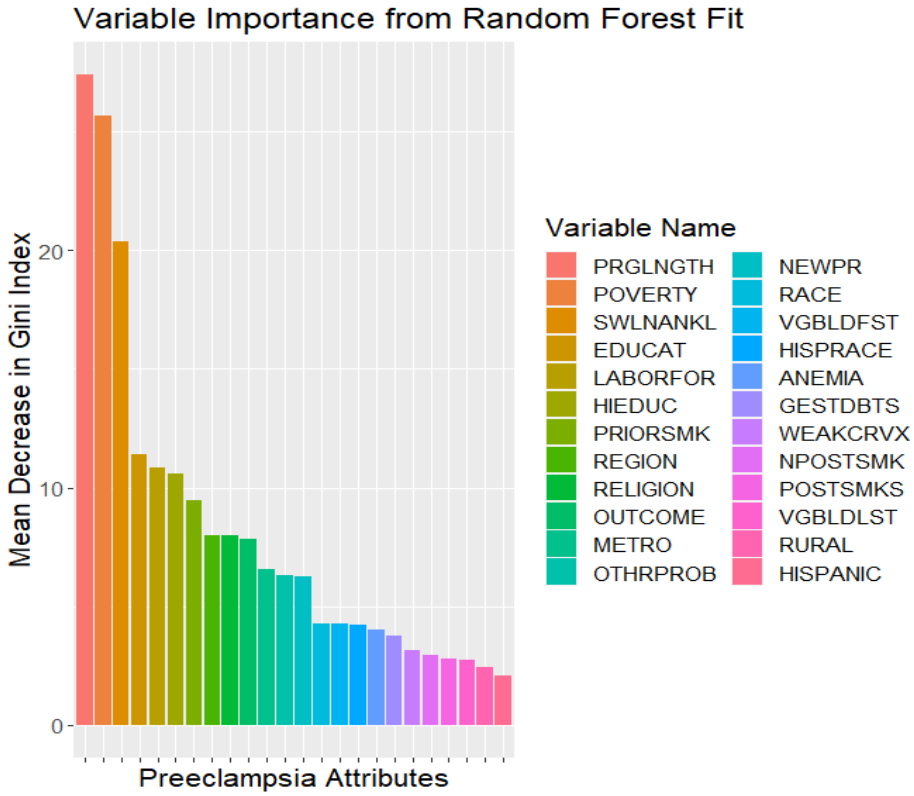
La interpretabilidad a nivel global de las predicciones se presenta por medio de una gráfica de barras, un gráfico de atributos, un árbol de decisión y un nomograma.

La interpretabilidad local se logra a través de LIME, en donde se explica, para instancias seleccionadas, cómo influyen las variables del conjunto de datos.

En primer lugar, se presenta en un gráfico (ver. Figura 4.17) las 25 variables que integran el conjunto de datos, ordenadas de acuerdo con su relevancia para la clasificación ordenadas por medio de MDGI (Mean Decrease Gini Index). MDGI mide la proporción de muestras clasificadas incorrectamente cuando la característica evaluada se elimina del

conjunto de datos Preeclampsia. Por lo que es un buen indicador de la relevancia general de las variables.

Figura 4.17 Variables más relevantes para el conjunto de datos Preeclampsia de acuerdo con el MDGI



Fuente: Elaboración propia.

La segunda forma de presentar los resultados de forma interpretable para el escenario de aplicación Preeclampsia, es un árbol de decisión. El árbol de decisión proporciona una visión global muy comprimida del comportamiento del modelo.

Las variables más importantes, de acuerdo con el índice GINI, son la duración del embarazo completado en semanas (PRGLNGTH), la pobreza, la retención de agua / edema en el embarazo (SWLNANKL), la educación (años de escolaridad completados) (EDUCAT), estado de la fuerza laboral (LABORFOR), el grado escolar más alto (HIEDUC).

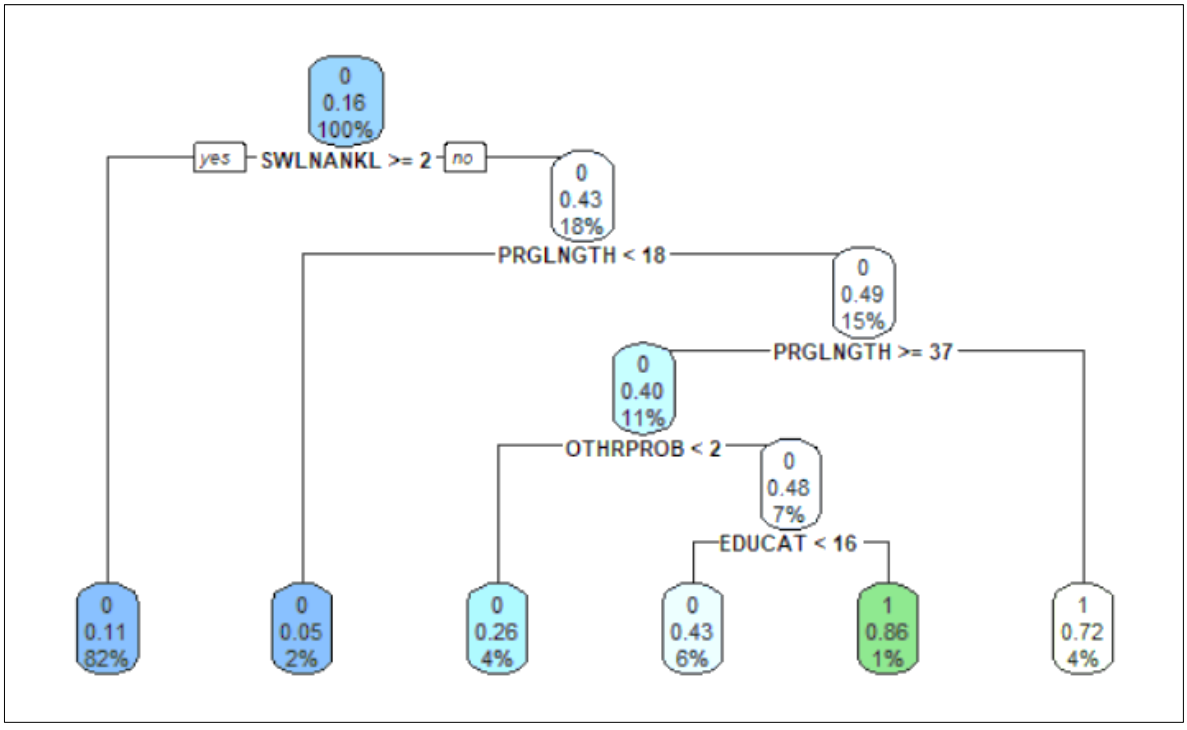
La estructura de árbol se adapta perfectamente para cubrir las interacciones entre las variables en los datos y la interpretación, además de proveer una visualización natural, por medio de los nodos y bordes.

A continuación, se presenta un árbol de decisión basado en tres variables del conjunto de datos (ver Figura 4.18).

En la primera línea se muestra el valor de la clase (0,1). En la segunda línea se muestran las probabilidades (0.16) de que la clase sea igual con "0" (no padece la enfermedad). En el primer nodo se pregunta si la variable SWLNANKL>=2. Si el valor de SWLNANKL no es mayor o igual que 2, entonces baja al nodo siguiente (nivel 2), en donde se muestra que el 18 por ciento de las personas cumplen con esa condición. Si PRGLNGTH es menor que 18, entonces se tiene el 15 por ciento de probabilidades de no padecer la enfermedad. En ese caso, se baja al siguiente nivel (nivel 4) por el lado derecho, en donde se cumple con la condición de PRLNGTH menor o igual que 37, lo que lleva al 4 por ciento de probabilidades de padecer la enfermedad (1).

En general, los dos números en la segunda fila de cada nodo indican la probabilidad de verse afectado por la variable, el número en la tercera fila representa el porcentaje de muestra cubierto por el nodo.

Figura 4.18 Árbol de decisión para el conjunto de datos Preeclampsia.



Fuente: Elaboración propia

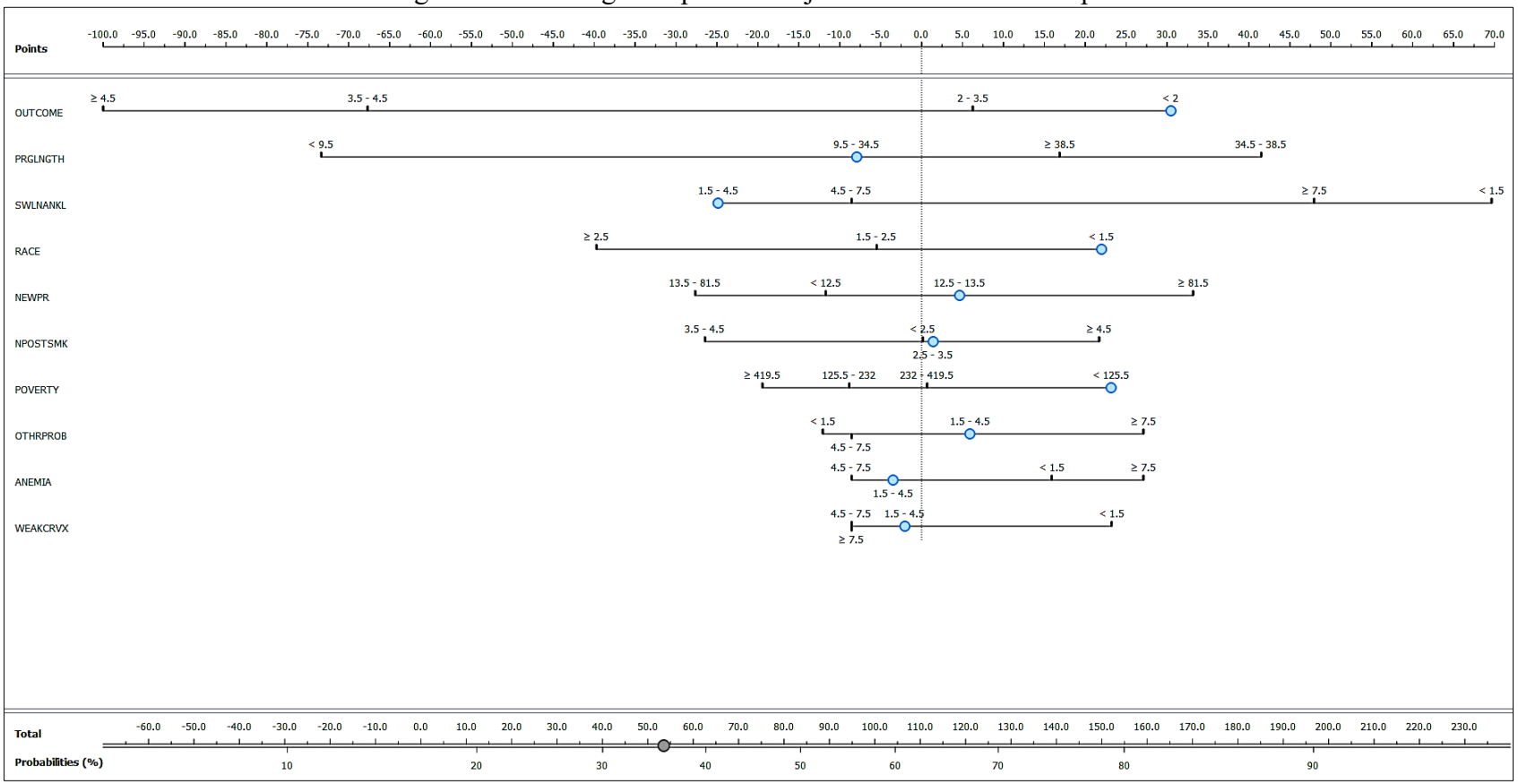
De esta manera los árboles de decisión muestran la forma en que se ha llegado a las inferencias que se presentan como resultado.

Otra explicación viable para las predicciones es un nomograma, que presenta las variables seleccionadas del conjunto de datos Preeclampsia y la probabilidad de padecer la enfermedad. En la Figura 4.19, se presenta el listado de diez variables, que permite valorar su influencia en la probabilidad de padecer la enfermedad. Esto se logra cambiando los valores de las variables de interés, alineando una recta con la escala de probabilidades que está en el límite inferior de la gráfica. Como ejemplo, los valores seleccionados para cada variable están indicados por medio de un círculo azul y las probabilidades de padecer la enfermedad están indicadas con un círculo gris en la escala inferior.

Como es evidente, la precisión de un nomograma está restringida por el número de variables enlistadas, así como por la alineación y percepción de los puntos que componen las escalas de valores para cada variable.

A pesar de sus limitaciones, los nomogramas muestran de manera sencilla cómo influyen los cambios en las variables en las predicciones de padecer la enfermedad.

Figura 4.19 Nomograma para el conjunto de datos Preeclampsia



Fuente: Elaboración propia.

Con el fin de proporcionar una visión local de los resultados obtenidos, se hace uso del modelo LIME para presentar algunos ejemplos del conjunto de datos Preeclampsia. Esto apoya al nivel de comprensión de las predicciones, pues permite validar la información generada.

En la Figura 4.20, se presentan algunos ejemplos, para explicar cuáles son los factores que resultan riesgosos para padecer Preeclampsia, así como algunos los que no lo son. Esto es, en términos de aprendizaje automático, las variables que apoyan la predicción y las que la contradicen. En la figura se muestra en color azul a las variables que soportan la predicción de padecer Preeclampsia, mientras que las variables que la contradicen se muestran en color rojo.

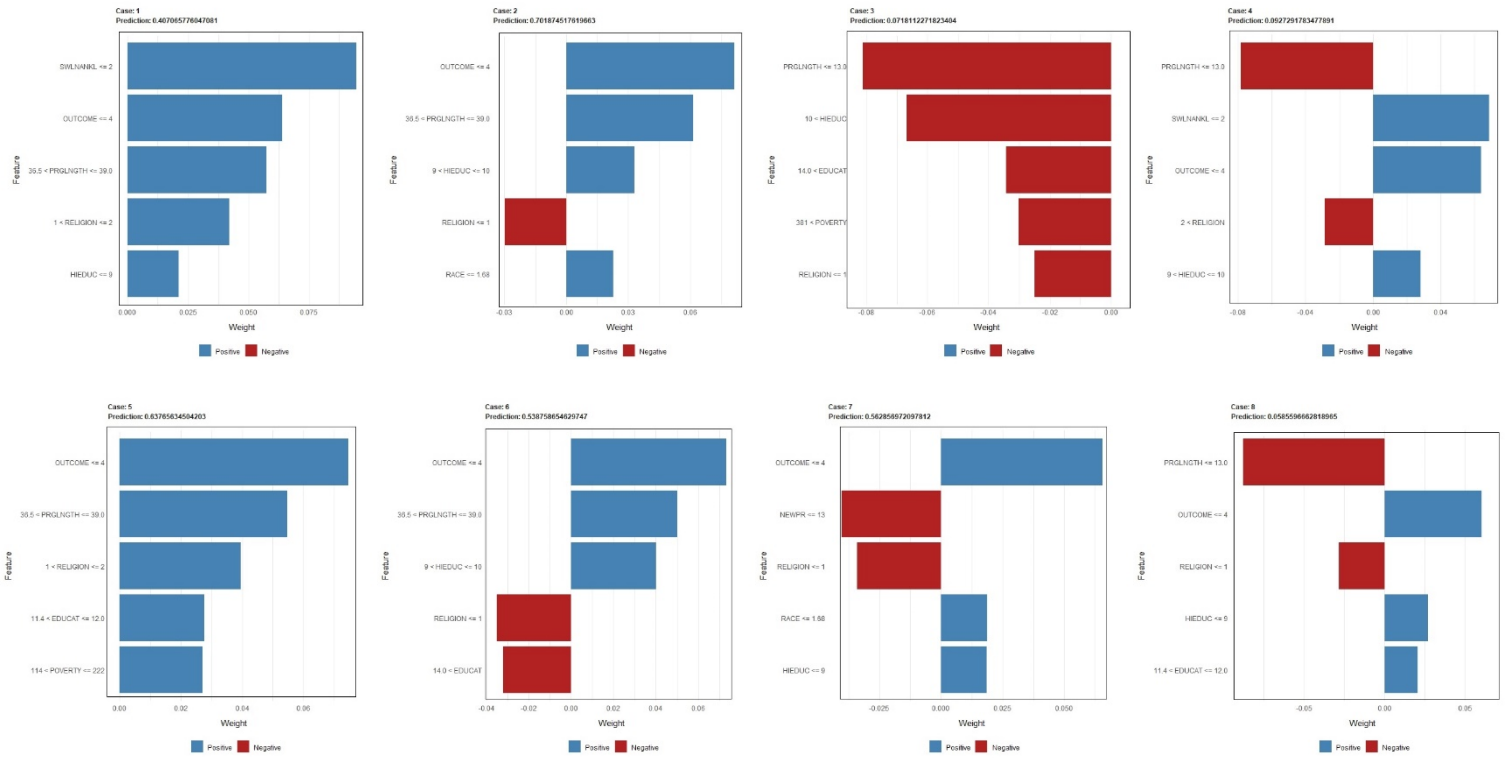
La gráfica se interpreta de la siguiente forma: para el primer elemento de la primera fila, se tiene el número de caso 1, con clase "1" (persona que padece Preeclampsia), las variables que son factores de riesgo son ausencia de toxemia (valores posibles 1= ausencia, 2=presencia, por convención se toma el valor menor o igual, como valor menor que 2) , duración del embarazo mayor que 36.5 y menor o igual que 39 semanas, resultado del embarazo menor o igual que 4 (producto vivo a nacer). el número de variables tomadas en cuenta para este diagrama es tres, por lo que no se reportan factores que contradigan la predicción para este caso.

Un ejemplo adicional es el segundo elemento de la segunda columna, con número de caso 4, con clase "0" (persona sin Preeclampsia), la variable que apoya la predicción es duración del embarazo menor o igual a 13 semanas. Mientras que las variables que contradicen la predicción son resultado del embarazo menor o igual que 4 (producto vivo al nacer) y ausencia de toxemia (valores posibles 1= ausencia, 2=presencia, por convención se toma el valor menor o igual, como valor menor que 2).

Esta representación gráfica, se elaboró tomando en cuenta tres variables que, si bien, no son suficientes para explicar por completo la forma en que se llega a las predicciones, permite lograr algún nivel de claridad. Si se agregan más variables el resultado será más informativo pero cada vez menos interpretable, en la medida que el número de variables aumente.

Por estas razones, es necesario establecer un equilibrio, de acuerdo con las necesidades de los usuarios finales de los resultados.

Figura 4.20 Interpretabilidad local para el conjunto de datos Preeclampsia



Fuente: Elaboración propia

4.3 Discusión

El aprendizaje automático ha tomado relevancia para el estudio de las enfermedades complejas (Noorbakhsh-Sabet, N., Zand, R., Zhang, Y., & Abedi, 2019). En la mayor parte no se aborda la problemática del dominio médico en su conjunto (Espinilla et al., 2017; Fergus et al., 2018; Fraccaro et al., 2015; Jiang et al., 2009; Kenny et al., 2005; Krishnaiah et al., 2015; Moreira et al., 2016; Neocleous et al., 2009; Spencer et al., 2011; Velikova et al., 2014). A continuación, se comparan los resultados obtenidos en trabajos analizados en el estado del arte los resultados logrados en este trabajo.

Con base en los trabajos previamente estudiados, se concluye que sólo en el caso de van der Schaar (Alaa & van der Schaar, 2018) se han aplicado varios de los pasos propuestos en esta metodología, para predecir la supervivencia a corto plazo de los pacientes con fibrosis quística. En este trabajo se preprocesan los datos, se aplican modelos para la clasificación, se menciona el problema del impacto del desbalanceo de las clases, y se presentan los resultados por medio de reglas extraídas de los modelos para explicar la predicción de supervivencia que separa a los pacientes que están realmente en riesgo de aquellos que no necesariamente necesitan un trasplante de pulmón a corto plazo. De forma que en el trabajo de van der Schaar no se incluyen la construcción de la base de datos, no se trata el problema del desbalanceo en las clases, se presentan los resultados de manera interpretable sólo a nivel general. En comparación con la metodología producto de esta investigación, el estudio de van der Schaar tiene un abordaje parcial del problema.

La metodología propuesta en esta investigación incluye un tratamiento integral del problema desde la recolección de datos hasta la presentación interpretable de los resultados. En seguida se contrastarán las investigaciones previas revisadas en el estado del arte, para cada escenario de aplicación.

Para el escenario de aplicación referente a DMRE, el problema de la identificación de factores de riesgo lo han tratado: Gold et al. (2006), Chen et al. (2007), Jiang et al. (2009), Spencer et al. (2011) y Fraccaro et al. (2015) sin tomar en cuenta el desbalanceo de los conjuntos de datos. A continuación, el problema lo abordaron Çelebiler et al. (2013), Krishnaiah S. et al. (2015), tomando en consideración el desbalanceo de los conjuntos de datos. Fraccaro et al. (2015) propuso también un sistema interpretable a través de árboles de decisión para el diagnóstico de DMRE.

En todos los casos mencionados, se ha abordado el problema de la identificación de factores de riesgo para DMRE considerando algunos aspectos de este. En ningún caso se ha abordado un procedimiento completo que implique la obtención de datos a partir del historial clínico de los sujetos, la toma de muestras de sangre, extracción de ADN, construcción de la base de datos, preprocesamiento de los datos, selección de variables más relevantes, balanceo del conjunto de datos, eliminación de instancias espurias, clasificación del conjunto de datos y la presentación interpretable de resultados.

Para preeclampsia, el segundo escenario de aplicación, diversos investigadores han estudiado el problema de la identificación de los factores de riesgo, entre ellos Neocleous et al., 2009, Cox et al., 2011, Tejera et al., 2011, Velikova et al., 2013, Mackenzie et al., 2016, Moreira et al., 2016, Fergus et al., 2018, Villa et al., 2017, Espinilla et al. (2017), Fergus et al., 2018, quienes no afrontan el problema en su conjunto. Kenny et al. (2005) identificaron factores de riesgo considerando el balanceo de las bases de datos y muestran árboles de decisión para mejorar la interpretabilidad de los resultados.

Los trabajos estudiados abordan los restos del aprendizaje máquina en el dominio del cuidado de la salud de manera parcial, es decir, no consideran la construcción de las bases de datos, preprocesamiento de los datos, selección de variables más relevantes, balanceo del conjunto de datos, eliminación de instancias espurias, clasificación del conjunto de datos y la presentación interpretable de resultados. En ningún trabajo se toma en cuenta un tratamiento integral que, finalmente, se presente como una metodología completa para el apoyo en la toma de decisiones.

En suma, los trabajos mostrados abordan el reto del estudio de los datos provenientes del dominio médico identificando algunos factores de riesgo, algunos tomando en cuenta el desbalanceo de los conjuntos de datos y la interpretabilidad sólo lo hacen al nivel de las predicciones.

La presente investigación propone una metodología que considera la obtención de datos a partir del historial clínico de los sujetos, la toma de muestras de sangre, extracción de ADN, construcción de la base de datos, preprocesamiento de los datos, selección de variables más relevantes, balanceo del conjunto de datos, eliminación de instancias espurias, clasificación del conjunto de datos y la presentación interpretable de resultados en sus tres niveles: el de los datos, el de los algoritmos y el de las predicciones.

El principal fin de los trabajos estudiados es identificar los factores de riesgo para las DMRE y PE, la utilidad de los modelos utilizados para la clasificación se cuantifica por medio de la precisión para identificar a las personas en riesgo de padecer la enfermedad. Muchos de los modelos desarrollados previamente han sido validados a través de la sensibilidad, especificidad.

Al aplicar algoritmos de aprendizaje automático sobre conjuntos de datos no balanceados, estos casi siempre producen clasificadores de alta Accuracy y especificidad, pero con baja sensibilidad (Chawla, 2010). El objetivo principal de aprender de conjuntos de datos balanceados es mejorar la sensibilidad sin dañar la Accuracy. Por esta razón, en este trabajo se cuantifica su efectividad una vez balanceados los conjuntos de datos.

Con el fin de dar claridad a la medida utilizada para evaluar los resultados del presente trabajo, después de balancear los conjuntos de datos, se utilizó Kappa pues es una métrica que compara el valor de Accuracy esperada contra la observada. El valor de Kappa de un modelo es directamente comparable con el de cualquier otro modelo utilizado para la clasificación. Para ambos conjuntos de datos, se encontró que *Random Forest* aplicado al ensamble de clasificadores resultó ser el que obtuvo mejores resultados, se obtuvo un valor de Kappa de 90.3 y 75.01 para DMRE y PE, respectivamente.

Adicionalmente, con el fin de comparar los resultados del presente trabajo con los previos, se presenta también la sensibilidad y especificidad para los ensambles de clasificadores seleccionados para cada escenario de aplicación. En donde la sensibilidad se puede expresar como la detección real de casos positivos y la especificidad como la detección real de los casos negativos. debido a la importancia de una predicción incorrecta en el área de la medicina, en donde un falso negativo implicaría que un individuo ha sido diagnosticado como sano, cuando en realidad padece la enfermedad. por esta razón estas métricas son las más utilizadas en el estado del arte.

A continuación, se compara el rendimiento alcanzado entre los trabajos analizados en el estado del arte con los obtenidos en esta investigación (ver Tabla 4.17). Para el escenario de aplicación DMRE, los valores de sensibilidad y especificidad de este trabajo son 93.73% y 95.56% respectivamente. Estos valores superan los obtenidos en los estudios revisados en este trabajo.

Para el otro escenario, el de Preeclampsia, los valores de sensibilidad y especificidad obtenidos en esta investigación son 72.09% y 89.50%, el valor de sensibilidad es más bajo que dos de los trabajos estudiados y mayo o igual que en los otros tres cotejados. En cuanto a la especificidad sólo el trabajo de Kenny et al. (2005) superó el valor obtenido, habiendo obtenido 98%. Los valores de sensibilidad y especificidad obtenidos en este escenario de aplicación ubican la sensibilidad en el conjunto de los trabajos con mejor puntuación y el valor de especificidad en el segundo lugar entre los mejor calificados.

Tabla 4.17 Comparación de los resultados obtenidos en varios estudios con los resultados del trabajo propuesto en esta investigación.

Degeneración Macular Relacionada con la Edad	
Estudio	**Resultados**
Metodología propuesta en este trabajo.	Sensibilidad: 93.73% Especificidad: 95.56%
Spencer et al. (2011)	Sensibilidad 77.0% Especificidad 74.1%
Jiang R. et al.(2009)	8.5% tasa de error de clasificación
Gold et al. (2006)	Sensibilidad: 58.37% Especificidad: 77.13%
Chen X. et al. (2007)	Tasa de falsos positivos < 5%
Çelebiler, A. et al.(2013)	Precisión: 83.3% (±4.7)
Fraccaro et al. (2015)	Rendimiento medio Caja Blanca: 92%, Caja Negra: 90%.
Krishnaiah S. et al. (2015)	Sensibilidad 79% Especificidad 69%
Preeclampsia	
Estudio	**Resultados**
Metodología propuesta en este trabajo.	Sensibilidad: 80% Especificidad: 89.50%
Kenny et al. (2005)	Sensibilidad: 100% Especificidad: 98%
Neocleous et al., (2009)	Sensibilidad: 50% Especificidad: 50%
Espinilla (2017)	Accuracy: 75.03% Alta interpretabilidad
Tejera et al., 2011	Sensibilidad: 80% Especificidad: 85-90%
Fergus et al., 2018hipótesis	Sensibilidad: 95% Especificidad: 87%

Conclusiones

Este trabajo presentó una metodología para apoyar al personal médico en el diagnóstico y pronóstico de algunas enfermedades complejas, con el fin de encontrar las posibles asociaciones entre ellas y sus factores de riesgo para afrontar los desafíos propios de los datos médicos y de interpretabilidad del modelo.

El propósito de este trabajo es integrar en una metodología que incluya: la creación de los conjuntos de datos, la detección de los factores de riesgo, el balanceo de los conjuntos de datos y la presentación de resultados de manera entendible para el personal médico.

Con base en lo antes expuesto, en los siguientes párrafos se contrastan las hipótesis planteadas con los resultados obtenidos.

Con referencia a la primera hipótesis (H1), los resultados muestran que a partir del aprendizaje automático se logró construir una herramienta de soporte para la toma de decisiones dirigida a expertos en el dominio médico. Esta se probó en los dos escenarios de aplicación del dominio médico: Degeneración Macular Relacionada con la Edad y Preeclampsia. Así, se presenta una metodología basada en las técnicas de aprendizaje automático para encontrar posibles asociaciones entre algunas enfermedades complejas y sus factores de riesgo con los siguientes elementos: construcción de la base de datos útiles para el pronóstico y/o diagnóstico de enfermedades complejas, balanceo de clases por medio de técnicas de sobremuestreo y eliminación de datos espurios; prueba y selección de algoritmos de clasificación adecuados para los datos; presentación de resultados en forma interpretable, considerando el nivel de datos, de algoritmos y de predicciones.

Con relación a la hipótesis (H2), referente al problema de la escasez de datos adecuados para la clasificación, se aplicaron técnicas de muestreo y limpieza de datos que mejoran el desempeño de los algoritmos de clasificación, se aplicó el algoritmo SMOTE, para sobre muestro y sub muestro de datos, que en el dominio médico es de especial relevancia debido a la dificultad para contar con los datos provenientes de personas enfermas y sanas, necesarios para tener clases balanceadas. La transformación y limpieza de los datos se ha

aplicado para las variables categóricas, que se presentan con frecuencia en los datos provenientes de historiales médicos, han mejorado el desempeño de los algoritmos de clasificación. Ambos procesos han probado influencia positiva por medio del valor de las métricas usadas para medir el desempeño de los clasificadores, que para los escenarios de aplicación de DMRE y Preeclampsia han alcanzado valores de Accuracy y Kappa de 95.18% y 90.29%; y de 90.99% y 81.49%, respectivamente. En consecuencia, la hipótesis se cumple.

Con relación a la hipótesis (H3) referente a la determinación de las posibles asociaciones entre los factores de riesgo y las enfermedades complejas, se observa en los resultados que esto se ha logrado con base en la técnica de eliminación de variables. Los resultados de los dos escenarios de aplicación se han contrastado con los factores presentes en la literatura de aplicación de aprendizaje máquina a la medicina. En particular, para el caso de DMRE se ha demostrado que los polimorfismos *rs203687* y *rs1329428* del gen CFH (*complement factor H*) son factores de riesgo para la población mexicana. Esto es relevante pues la componente genética del grupo poblacional mexicano pertenece al conjunto de etnias que no habían sido estudiadas para esta enfermedad. En consecuencia, se concluye que la hipótesis es verdadera.

La hipótesis relacionada con la presentación comprensible de los resultados obtenidos (H4) se cumple, pues se muestran los resultados generales y a nivel individual por medio de técnicas de interpretabilidad. Se han contemplado los tres niveles de interpretabilidad: se explica el tipo de datos y la relevancia de ellos, se presentan los algoritmos que se han utilizado para elaborar el análisis, y se explica la forma en que se llegó a las predicciones. La forma de llegar a las predicciones se presenta a nivel general por medio de árboles de decisión, reglas de inferencia y nomogramas; a nivel individual, los resultados obtenidos se muestran por medio de diagramas basados en LIME (*Local Interpretable Model-agnostic Explanations)*. De esta forma, los expertos médicos pueden verificar cómo se infirieron los resultados.

De acuerdo con lo establecido por Le Cun et al. (2015), en años recientes, los sistemas de inteligencia artificial y aprendizaje automático han logrado un rendimiento en muchas tareas que se consideraba computacionalmente inalcanzable. Sin embargo, el gran reto para la adopción de la inteligencia artificial en el ámbito médico ha sido presentar los resultados con generalidad suficiente y explicación de los resultados para los casos individuales de

forma clara. Como se puede observar, en este trabajo se han cubierto el enfoque de precisión en los resultados y el de presentación accesible para los expertos del área médica.

La aportación de este trabajo se centra en proponer una metodología que permita obtener resultados con un nivel alto de precisión en los resultados y accesibles para los posibles usuarios de estos. Este trabajo contribuye con una herramienta para apoyar a los especialistas médicos en la prevención y diagnóstico temprano de las enfermedades complejas, que aquejan a millones de personas en México y en el mundo. Esto puede permitir el ahorro en el gasto en el tratamiento de enfermedades para las familias y los gobiernos. Así como, la consecuente mejoría en la calidad de vida de las personas.

Los resultados obtenidos, medidos por medio de sensibilidad y especificidad, ubican al caso de DMRE como el mejor evaluado entre los estudios analizados en este trabajo. Para el caso de Preeclampsia, la sensibilidad obtenida sitúa al estudio en el conjunto de los mejores evaluados y la especificidad como el segundo mejor evaluado.

Una limitación para realizar este trabajo ha sido la obtención de los datos por la escasez y dificultad para obtener datos útiles. Para reafirmar los hallazgos encontrados con referencia a los factores de riesgo y aplicarlos en el dominio médico, es deseable validarlos con mayor número de individuos para ambos escenarios de aplicación. En este sentido, una observación relevante de Poursabzi-Sangdeh et al. (2018), es que los tomadores de decisiones deberán considerar un escepticismo saludable para verificar empíricamente los modelos interpretables antes de aplicar los resultados de un modelo.

En trabajos futuros se puede trabajar en la definición de métricas particulares para los conjuntos de datos con las variables particulares de los provenientes del ámbito médico, en particular de las enfermedades complejas. También es deseable considerar técnicas de balanceo que incluyan métricas que vayan más allá de la tasa de desbalanceo, como el discriminante máximo de Fisher, así como distribuir los datos en grupos por medio de técnicas de "*clustering"* antes de balancearlos con el fin de lograr un balanceo más eficiente y mejorar el desempeño de los clasificadores.

Esta tesis ha permitido realizar los trabajos que se mencionan a continuación:

[1] A. Martínez-Velasco et al., "Assessment of CFH and HTRA1 polymorphisms in age-related macular degeneration using classic and machine-learning approaches," *Ophthalmic Genet.,* vol. 41, no. 6, pp. 539–547, 2020, doi: 10.1080/13816810.2020.1804945.

[2] A. Martínez-Velasco, L. Martínez-Villaseñor, L. Miralles-Pechuán, A. C. Perez-Ortiz, J. C. Zenteno, and F. J. Estrada-Mena, "The relevance of cataract as a risk factor for age-related macular degeneration: A machine learning approach," *Appl. Sci.*, vol. 9, no. 24, 2019, doi: 10.3390/app9245550.

[3] A. Martínez-Velasco, A. C. Perez-Ortiz, J. C. Zenteno, and F. J. Estrada-Mena L. Martínez-Villaseñor, "CFH and HTRA1 genes associated with AMD in Mexican population," *Invest. Ophthalmol. Vis. Sci.*, vol. 58, no. 8, p. 2268, 2017.

[4] A. Martínez-Velasco and L. Martínez-Villaseñor, "A Survey of Machine Learning Approaches for Age Related Macular Degeneration Diagnosis and Prediction," in *MICAI 2017*, 2017, vol. 10632 LNAI, pp. 257–266, doi: 10.1007/978-3-030-02837-4_21.

[5] A. Martinez-Velasco, L. Martínez Villaseñor, and L. Miralles, "Machine Learning Approach for Pre-Eclampsia Risk Factors Association," in *Goodtechs '18 Proceedings of the 4th EAI International Conference on Smart Objects and Technologies for Social Good*, 2018, no. January 2019, pp. 232–237, doi: 10.1145/3284869.3284912.

[6] A. Martínez-Velasco *et al.*, "Machine Learning Method to Establish the Connection Between Age Related Macular Degeneration and Some Genetic Variations," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10, no. 4, pp. 28–39, 2017, doi: 10.1007/978-3-319-48799-1.

Glosario

CFH	Gen que proporciona instrucciones para producir una proteína llamada factor H del complemento.
DMRE	Degeneración Macular relacionada con la edad.
GBM	Generalized Boosted Regression Models.
GLM	Generalización flexible de la regresión lineal.
HTRA1	Gen proporciona instrucciones para producir una proteína que se encuentra en muchos de los órganos y tejidos del cuerpo.
KNN	K nearest neighborhood
LDA	Análisis Discriminante Lineal (Linear Discrimiant Analysis).
LIME	Local Interpretable Model-agnostic Explanations
MDGI	Mean Decrease Gini Index.
PE	Preeclampsia.
RFE	(Recursive Feature Elimination).
ROC	Receiver Operating Curve.
RPART	Técnica estadística de análisis multivariante cuyo objetivo es construir árboles de decisión que modelen la influencia de una serie de variables explicativas sobre la variable objetivo de un estudio estadístico.
SMOTE	Synthetic Minority Over-sampling Technique.
SVM	Support Vector Machine.

Referencias

Alaa, A. M., & van der Schaar, M. (2018). Prognostication and Risk Factors for Cystic Fibrosis via Automated Machine Learning. *Scientific Reports*, *8*(1), 11242. https://doi.org/10.1038/s41598-018-29523-2

Alanis Tamez, M. D. (2018). *Prediagnóstico de enfermedades crónicas mediante algoritmos de cómputo inteligente* [CIC, IPN]. https://tesis.ipn.mx/bitstream/handle/123456789/26201/T1948.pdf?sequence=1&isAllowed=y

Avni, E., & Snir, S. (2019). A new quartet-based statistical method for comparing sets of gene trees is developed using a generalized hoeffding inequality. *Journal of Computational Biology*, *26*(1), 27–37. https://doi.org/10.1089/cmb.2018.0129

Bach, M., Werner, A., Żywiec, J., & Pluskiewicz, W. (2017). The study of under- and over-sampling methods' utility in analysis of highly imbalanced data on osteoporosis. *Information Sciences*, *384*, 174–190. https://doi.org/10.1016/j.ins.2016.09.038

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. F., & Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: An overview. In *Bioinformatics* (Vol. 16, Issue 5, pp. 412–424). https://doi.org/10.1093/bioinformatics/16.5.412

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, *6*(1), 20. https://doi.org/10.1145/1007730.1007735

Berrar, D. (2019). Performance Measures for Binary Classification. In *Encyclopedia of Bioinformatics and Computational Biology* (pp. 546–560). Academic Press. https://doi.org/10.1016/b978-0-12-809633-8.20351-8

Bica, I., Alaa, A. M., Lambert, C., & van der Schaar, M. (2020). From real-world patient data to individualized treatment effects using machine learning: Current and future methods to address underlying challenges. *Clinical Pharmacology & Therapeutics*.

Blagus, R., & Lusa, L. (2013). SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics*, *14*. https://doi.org/10.1186/1471-2105-14-106

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brown, G. (2010). *Ensemble Learning Motivation and Background*. Springer Press.

Brunotto, M., & Zárate, A. M. (2012). Modelos predictivos para enfermedades complejas. *Revista Facultad de Ciencias Médicas Univ.NAcional Córdoba.*, *69*(1), 33–41. http://www.revista2.fcm.unc.edu.ar/2012.69.1/Revision/revision33-41.pdfhttp://oca.unc.edu.ar/

Cacheiro Martínez, P., Ordovás, J. M., & Corella, D. (2011). *Métodos de selección de variables en estudios de asociación genética. Aplicación a un estudio de genes candidatos en Enfermedad de Parkinson.*

Camaré, L. J. M. (2008). *Aprendizaje Automático a partir de Conjuntos de Datos No Balanceados y su Aplicación en el Diagnóstico y Pronóstico Médico*.

Caparrini, F. S., & de J. Pérez Jiménez, M. (2002). *Verificación de programas en modelos de computación no convencionales*. Universidad de Sevilla.

Castaneda, C., Nalley, K., Mannion, C., Bhattacharyya, P., Blake, P., Pecora, A., Goy, A., & Suh, K. S. (2015). Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine. *Journal of Clinical Bioinformatics*, *5*(1), 4. https://doi.org/10.1186/s13336-015-0019-3

Çelebiler, A., Şeker, H., Yuksel, B., Bilgilil, A., & Karaca, M. B. (2013). Discovery of the connection among age-related macular degeneration, MTHFR C677T and PAI 1 4G/5G gene polymorphisms, and body mass index by means of Bayesian inference methods. *Turkish Journal of Electrical Engineering & Computer Sciences*, *21*(SUPPL. 1), 2062–2078. https://doi.org/10.3906/elk-1111-21

Centro del Conocimiento Bioético. (2015). Comisión Nacional de Bioética :: México. In *Bioética*. http://www.conbioetica-mexico.salud.gob.mx/interior/temasgeneral/consentimiento_informado.html

Cerón-Mireles, P., Harlow, S. D., Sánchez-Carrillo, C. I., & Núñez, R. M. (2001). Risk factors for pre-eclampsia/eclampsia among working women in Mexico City. *Paediatric and Perinatal Epidemiology*, *15*(1), 40–46.

http://www.ncbi.nlm.nih.gov/pubmed/11237114

Chakravarthy, U., Wong, T. Y., Fletcher, A., Piault, E., Evans, C., Zlateva, G., Buggage, R., Pleil, A., & Mitchell, P. (2010). Clinical risk factors for age-related macular degeneration: a systematic review and meta-analysis. *BMC Ophthalmology*, *10*(1), 1–13. https://doi.org/10.1186/1471-2415-10-31

Chawla, N. (2002). *SMOTE-N*. Carnegie Mellon University. School of Computer Science Web Site. https://www.cs.cmu.edu/afs/cs/project/jair/pub/volume16/chawla02a-html/node16.html

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Chawla, N. V., Lazarevic, A., Hall, L. O., & Bowyer, K. W. (2003). SMOTEBoost: Improving prediction of the minority class in boosting. *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)*, *2838*, 107–119. https://doi.org/10.1007/978-3-540-39804-2_12

Chawla, N. V. (2010). Data Mining for Imbalanced Datasets: An Overview Nitesh. *Data Mining and Knowledge Discovery Handbook*. https://doi.org/10.1007/978-0-387-09823-4

Chen, X., Liu, C.-T., Zhang, M., & Zhang, H. (2007). A forest-based approach to identifying gene and gene gene interactions. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(49), 19199–19203. https://doi.org/10.1073/pnas.0709868104

Corso, C. L., & Gibellini, F. (2011). *Aplicación de redes bayesianas usando Weka.* 939–948.

Cox, B., Sharma, P., Evangelou, A. I., Whiteley, K., Ignatchenko, V., Ignatchenko, A., Baczyk, D., Czikk, M., Kingdom, J., Rossant, J., Gramolini, A. O., Adamson, S. L., & Kislinger, T. (2011). Translational Analysis of Mouse and Human Placental Protein and mRNA Reveals Distinct Molecular Pathologies in Human Preeclampsia. *Molecular & Cellular Proteomics*, *10*(12), M111.012526. https://doi.org/10.1074/mcp.M111.012526

Craig, J. (2008). Complex Diseases: Research and Applications. *Nature Education*, *1*(1), 184. http://faculty.dbmi.pitt.edu/cosbbi/cosbbi2014/CoSBBI_Reading_ComplexDiseases.pdf

Dasgupta, A., & Sun, Y. (2011). Brief review of regression-based and machine learning methods in genetic epidemiology: the Genetic Analysis Workshop 17 experience. *Genetic Epidemiology*, *35*(Suppl 1), 1–13. https://doi.org/10.1002/gepi.20642.Brief

Denil, M., & Trappenberg, T. (2010). Overlap versus imbalance. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *6085 LNAI*, 220–231. https://doi.org/10.1007/978-3-642-13059-5_22

DeWan, A., Liu, M., Hartman, S., Zhang, S. S.-M., Liu, D. T. L., Zhao, C., Tam, P. O. S., Chan, W. M., Lam, D. S. C., Snyder, M., Barnstable, C., Pang, C. P., & Hoh, J. (2006). HTRA1 promoter polymorphism in wet age-related macular degeneration. *Science (New York, N.Y.)*, *314*(5801), 989–992. https://doi.org/10.1126/science.1133807

Dietterich, T. G. (2000). Experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. *Machine Learning*, *40*(2), 139–157. https://doi.org/10.1023/A:1007607513941

Díez-Pastor, J., Rodríguez, J. J., García-Osorio, C., & Kuncheva, L. (2015). Diversity techniques improve the performance of the best imbalance learning ensembles. *Information Sciences*, *325*, 98–117. https://doi.org/10.1016/j.ins.2015.07.025

Doran, D., Schulz, S., & Besold, T. R. (2018). What does explainable AI really mean? A new conceptualization of perspectives. *CEUR Workshop Proceedings*, *2071*. http://amueller.github.

Doshi-Velez, F., & Kim, B. (2017). *Towards A Rigorous Science of Interpretable Machine Learning*. *Ml*, 1–13. https://doi.org/10.1016/j.bbrc.2004.04.155

Dosilovic, F. K., Brcic, M., & Hlupic, N. (2018). Explainable artificial intelligence: A survey. *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2018 - Proceedings*, 210–215. https://doi.org/10.23919/MIPRO.2018.8400040

Duckitt, K., & Harrington, D. (2005). Risk factors for pre-eclampsia at antenatal booking: Systematic review of controlled studies. *British Medical Journal*, *330*(7491), 565–567. https://doi.org/10.1136/bmj.38380.674340.E0

Edwards, L., & Veale, M. (2017). Slave to the Algorithm? Why a Right to Explanationn is Probably Not the Remedy You are Looking for. *SSRN Electronic Journal*, *16*. https://doi.org/10.2139/ssrn.2972855

Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H., & Nadeau, J. H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics*, *11*(6), 446–450. https://doi.org/10.1038/nrg2809

Espinilla, M., Medina, J., García-Fernández, Á.-L., Campaña, S., & Londoño, J. (2017). Fuzzy Intelligent System for Patients with Preeclampsia in Wearable Devices. *Mobile Information Systems*, *2017*, 1–10. https://doi.org/10.1155/2017/7838464

Fergus, P., Montanez, C. C., Abdulaimma, B., Lisboa, P., & Chalmers, C. (2018). *Utilising Deep Learning and Genome Wide Association Studies for Epistatic-Driven Preterm Birth Classification in African-American Women*. 1–11. http://arxiv.org/abs/1801.02977

Fernández, A., García, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. In *Journal of Artificial Intelligence Research* (Vol. 61, pp. 863–905). https://doi.org/10.1613/jair.1.11192

Fernández, A., López, V., Galar, M., Del Jesus, M. J., & Herrera, F. (2013). Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches. *Knowledge-Based Systems*, *42*, 97–110. https://doi.org/10.1016/j.knosys.2013.01.018

Fraccaro, P., Nicolo, M., Bonetto, M., Giacomini, M., Weller, P., Traverso, C. E., Prosperi, M., OSullivan, D., & O 'sullivan, D. (2015). Combining macula clinical signs and patient characteristics for age-related macular degeneration diagnosis: a machine learning approach. *BMC Ophthalmology*, *15*, 10. https://doi.org/10.1186/1471-2415-15-10

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2012). A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. In *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews* (Vol.

42, Issue 4, pp. 463–484). https://doi.org/10.1109/TSMCC.2011.2161285

García, V., Sánchez, J. S., & Mollineda, R. A. (2012). On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. *Knowledge-Based Systems*, *25*(1), 13–21. https://doi.org/10.1016/j.knosys.2011.06.013

Gold, B., Merriam, J. C. J. E., Zernant, J., Hancox, L. S., Taiber, A. J., Gehrs, K., Cramer, K., Neel, J., Bergeron, J., Barile, G. R., Smith, R. T., Hageman, G. S., Dean, M., Allikmets, R., Chang, S., Yannuzzi, L. A., Merriam, J. C. J. E., Barbazetto, I., Lerner, L. E., … Stockman, H. (2006). Variation in factor B (BF) and complement component 2 (C2) genes is associated with age-related macular degeneration. *Nature Genetics*, *38*(4), 458–462. https://doi.org/10.1038/ng1750

Goldberg, J., Flowerdew, G., Smith, E., Brody, J. A., & Tso, M. O. (1988). Factors associated with age-related macular degeneration. An analysis of data from the first National Health and Nutrition Examination Survey. *American Journal of Epidemiology*, *128*.

González, A. (2014). Conceptos básicos de Machine Learning. *CleverTask It to Business Solutions*. https://cleverdata.io/que-es-machine-learning-big-data/

Guo, H., & Viktor, H. L. (2004). Learning from imbalanced data sets with boosting and data generation. *ACM SIGKDD Explorations Newsletter*, *6*(1), 30. https://doi.org/10.1145/1007730.1007736

Guo, X., Yin, Y., Dong, C., Yang, G., & Zhou, G. (2008). On the class imbalance problem. *Proceedings - 4th International Conference on Natural Computation, ICNC 2008*, *4*, 192–201. https://doi.org/10.1109/ICNC.2008.871

Hindorff, L., MacArthur, J., Junkins HA, Hall PN, Klemm AK, & Manolio TA. (2014). *Catalog of Published Genome-Wide Association Studies - National Human Genome Research Institute (NHGRI)*. https://www.genome.gov/gwastudies/

Iniesta, R., Guinó, E., & Moreno, V. (2005). Análisis estadístico de polimorfismos genéticos en estudios epidemiológicos. *Gaceta Sanitaria*, *19*(4), 333–341. https://doi.org/10.1157/13078029

Inter-university Consortium for Political and Social Research. (2008). *Public Private Ventures. Evaluation of Children's Futures: Improving Health*

and Development Outcomes for Children in Trenton, New Jersey, 2001-2005. https://doi.org/doi.org/10.3886/ICPSR21640.v1

Jain, A., Ratnoo, S., & Kumar, D. (2018). Performance comparison of classification algorithms for medical diagnosis. *Pertanika Journal of Science and Technology*, *26*(2), 729–748.

Janecek, A., Gansterer, W. N. W., Demel, M., & Ecker, G. (2008). On the Relationship Between Feature Selection and Classification Accuracy. *Fsdm*, *4*, 90–105.

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis*, *6*(5), 429–449. https://doi.org/10.3233/ida-2002-6504

Jiang, R., Tang, W., Wu, X., & Fu, W. (2009). A random forest approach to the detection of epistatic interactions in case-control studies. *BMC Bioinformatics*, *10 Suppl 1*(SUPPL. 1), S65. https://doi.org/10.1186/1471-2105-10-S1-S65

Jimenez-Corona, A., & Graue-Hernandez, E. O. (2018). Global prevalence and years lived with disability (YLDs) due to vision loss in Mexico in 2016. *Investigative Ophthalmology and Visual Science*, *59*(9). https://iovs.arvojournals.org/article.aspx?articleid=2691791

Kenny, L. C., Dunn, W. B., Ellis, D. I., Myers, J., Baker, P. N., & Kell, D. B. (2005). Novel biomarkers for pre-eclampsia detected using metabolomics and machine learning. *Metabolomics*, *1*(3), 227–234. https://doi.org/10.1007/s11306-005-0003-1

Khalid, S., Khalil, T., & Nasreen, S. (2014). A Survey of Feature Selection and Feature Extraction Techniques in Machine Learning. *Science and Information Conference*, 372–378. https://doi.org/10.1109/SAI.2014.6918213

Khan, S. H., Hayat, M., Bennamoun, M., Sohel, F. A., & Togneri, R. (2018). Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(8), 3573–3587. https://doi.org/10.1109/TNNLS.2017.2732482

Kotsiantis, S. B., & Pintelas, P. E. (2004). Recent Advances in Clustering: A Brief Survey. *WSEAS Transactions on Information Science and Applications*, *1*(1), 73–81.

Kotsiantis, S., Patriarcheas, K., & Xenos, M. (2010). A combinational

incremental ensemble of classifiers as a technique for predicting students' performance in distance education. *Knowledge-Based Systems*, *23*(6), 529–535. https://doi.org/10.1016/j.knosys.2010.03.010

Kotsiantis, S. S. B. S. (2007). Supervised machine learning: a review of classification techniques. *Informatica (Ljubljana)*, *31*(3), 249–268. https://datajobs.com/data-science-repo/Supervised-Learning-[SB-Kotsiantis].pdf

Kovács, G. (2019). Smote-variants: A python implementation of 85 minority oversampling techniques. *Neurocomputing*, *366*, 352–354. https://doi.org/10.1016/j.neucom.2019.06.100

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, *5*(4), 221–232. https://doi.org/10.1007/s13748-016-0094-0

Krishnaiah, S., Surampudi, B., & Keeffe, J. (2015). Modeling the risk of age-related macular degeneration and its predictive comparisons in a population in South India. *International Journal of Community Medicine and Public Health*, *2*(2), 137. https://doi.org/10.5455/2394-6040.ijcmph20150514

Ladha, L., & Deepa, T. (2011). Feature Selection Methods And Algorithms. *International Journal on Computer Science and Engineering*, *3*(5), 1787–1797. http://journals.indexcopernicus.com/abstract.php?icid=945099

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement of categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Larrañaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., Lozano, J. A., Armañanzas, R., Santafé, G., Pérez, A., & Robles, V. (2006). Machine learning in bioinformatics. *Briefings in Bioinformatics*, *7*(1), 86–112. http://www.ncbi.nlm.nih.gov/pubmed/16761367

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*.

Liu, X. Y., & Zhou, Z. H. (2006). The influence of class imbalance on cost-sensitive learning: An empirical study. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 970–974. https://doi.org/10.1109/ICDM.2006.158

Londoño Agudelo, E. (2017). Chronic diseases and the unavoidable transformation of health systems in Latin America. *Revista Cubana de*

Salud Publica, *43*(1), 68–74. http://scielo.sld.cu

López Takeyas, B. (2012). *Intoduccion a la Inteligencia Artificial*. *1956*, 1–20. http://www.cs.bham.ac.uk/~rmp/slide_book/slide

Luque, A., Carrasco, A., Martín, A., & de las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, *91*, 216–231. https://doi.org/10.1016/j.patcog.2019.02.023

Maddox, T. M., Rumsfeld, J. S., & Payne, P. R. (2019). Questions for artificial intelligence in health care. *JAMA - Journal of the American Medical Association*.

Martínez-Velasco, A., & Martínez-Villaseñor, L. (2017). A Survey of Machine Learning Approaches for Age Related Macular Degeneration Diagnosis and Prediction. *MICAI 2017*, *10632 LNAI*, 257–266. https://doi.org/10.1007/978-3-030-02837-4_21

Martinez-Velasco, A., Martínez Villaseñor, L., & Miralles, L. (2018). Machine Learning Approach for Pre-Eclampsia Risk Factors Association. In B. Guidi (Ed.), *Goodtechs '18 Proceedings of the 4th EAI International Conference on Smart Objects and Technologies for Social Good* (Issue January 2019, pp. 232–237). https://doi.org/10.1145/3284869.3284912

McCarthy, K., Zabar, B., & Weiss, G. (2005). Does cost-sensitive learning beat sampling for classifying rare classes? *Proceedings of the 1st International Workshop on Utility-Based Data Mining, UBDM '05*, *June*, 69–77. https://doi.org/10.1145/1089827.1089836

Mehta, R., Tech, M., Bhatt, N., & Ganatra, A. (2016). A Survey on Data Mining Technologies for Decision Support System of Maternal Care Domain. *International Journal of Computer Applications*, *138*(10), 975–8887. https://doi.org/10.5120/ijca2016908965

Mena, L. J., Orozco, E. E., Felix, V. G., Ostos, R., Melgarejo, J., & Maestre, G. E. (2012). Machine learning approach to extract diagnostic and prognostic thresholds: Application in prognosis of cardiovascular mortality. *Computational and Mathematical Methods in Medicine*, *2012*. https://doi.org/10.1155/2012/750151

Menardi, G., & Torelli, N. (2014). Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery*, *28*(1), 92–122. https://doi.org/10.1007/s10618-012-0295-5

Mohammed, A. J., Hassan, M. M., & Kadir, D. H. (2020). Improving classification performance for a novel imbalanced medical dataset using smote method. *International Journal of Advanced Trends in Computer Science and Engineering*, *9*(3), 3161–3172. https://doi.org/10.30534/ijatcse/2020/104932020

Montavon, G., Samek, W., & Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing: A Review Journal*, *73*, 1–15. https://doi.org/10.1016/j.dsp.2017.10.011

Moor, J. (2006). The Dartmouth College Artificial Intelligence Conference: The Next Fifty Years. *AI Magazine*, *27*(4), 87–87. https://doi.org/10.1609/AIMAG.V27I4.1911

Moreira, M. W. L., Rodrigues, J. J. P. C., Oliveira, A. M. B., Ramos, R. F., & Saleem, K. (2016). A preeclampsia diagnosis approach using Bayesian networks. *2016 IEEE International Conference on Communications (ICC)*, 1–5. https://doi.org/10.1109/ICC.2016.7510893

Najman, J. M., Morrison, J., Williams, G. M., Keeping, J. D., & Andersen, M. J. (1989). Unemployment and reproductive outcome. An Australian study. *British Journal of Obstetrics and Gynaecology*, *96*(3), 308–313. http://www.ncbi.nlm.nih.gov/pubmed/2713289

National Institutes of Health, N. (2014). Diccionario de cáncer - National Cancer Institute. In *US Government* (p. 1). https://www.cancer.gov/espanol/publicaciones/diccionario

Neocleous, C. K., Anastasopoulos, P., Nikolaides, K. H., Schizas, C. N., & Neokleous, K. C. (2009). Neural networks to estimate the risk for preeclampsia occurrence. *Proceedings of the International Joint Conference on Neural Networks*, 2221–2225. https://doi.org/10.1109/IJCNN.2009.5178820

Noorbakhsh-Sabet, N., Zand, R., Zhang, Y., & Abedi, V. (2019). Artificial intelligence transforms the future of health care. *The American Journal of Medicine*, *795–801*. https://doi.org/10.1016/j.amjmed.2019.01.017

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. *New England Journal of Medicine*, *375*(13), 1216–1219. https://doi.org/10.1056/nejmp1606181

Peek, N., Combi, C., Marin, R., & Bellazzi, R. (2015). Thirty years of artificial intelligence in medicine (AIME) conferences: A review of

research themes. In *Artificial Intelligence in Medicine* (Vol. 65, Issue 1, pp. 61–73). https://doi.org/10.1016/j.artmed.2015.07.003

R. Duda, P. E. H., & Stok, D. G. (2001). *Pattern Classification* (JOHN WILEY & SONS INC (ed.); 2nd.).

Ramyachitra, D., & Manikandan, P. (2014). Imbalanced Dataset Classification and Solutions: a Review. *International Journal of Computing and Business Research (IJCBR) ISSN (Online*, *5*(4), 2229–6166.

Reddy, S., Fox, J., & Purohit, M. P. (2019). Artificial intelligence-enabled healthcare delivery. *Journal of the Royal Society of Medicine*, *112*(1), 22–28.

Ren, F., Cao, P., Li, W., Zhao, D., & Zaiane, O. (2017). Ensemble based adaptive over-sampling method for imbalanced data learning in computer aided detection of microaneurysm. *Computerized Medical Imaging and Graphics*, *55*, 54–67. https://doi.org/10.1016/j.compmedimag.2016.07.011

Riancho, J. A. (2012). Enfermedades complejas y análisis genéticos por el método GWAS. Ventajas y limitaciones. *Reumatología Clínica*, *8*(2), 56–57. https://doi.org/10.1016/j.reuma.2011.07.005

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. 1135–1144. https://doi.org/10.18653/v1/N16-3020

Rivera, W. A., & Xanthopoulos, P. (2016). A priori synthetic over-sampling methods for increasing classification sensitivity in imbalanced data sets. *Expert Systems with Applications*, *66*, 124–135. https://doi.org/10.1016/j.eswa.2016.09.010

Roberts, J. M., Druzin, M., August, P. A., Gaiser, R. R., Bakris, G., Granger, J. P., Barton, J. R., Jeyabalan, A., Bernstein, I. a, Johnson, D. D., Karamanchi, S. A., Spong, C. Y., Lindheiner, M. D., Tsingas, E., Owens, M. Y., Martin Jr, J. N., Saade, G. R., & Sibai, B. M. (2012). ACOG Guidelines: Hypertension in pregnancy. In *American College of Obstetricians and Gynecologists*. https://doi.org/doi: 10.1097/01.AOG.0000437382.03963.88

Rodríguez Torres, F. (2017). *Smote-d, una versión determinista de smote.* INAOE.

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more

informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*, *10*(3), 1–21. https://doi.org/10.1371/journal.pone.0118432

Sedgwick, P. (2014). Nested case-control studies: Advantages and disadvantages. In *BMJ (Online)* (Vol. 348). https://doi.org/10.1136/bmj.g1532

Sing, C. F., Haviland, M. B., Templeton, A. R., Zerba, K. E., & Reilly, S. L. (1992). Biological complexity and strategies for finding DNA variations responsible for inter-individual variation in risk of a common chronic disease, coronary artery disease. *Annals of Medicine*, *24*(6), 539–545. https://doi.org/10.3109/07853899209167008

Singh Malik, J., Goyal, P., & Sharma, M. K. (2007). A Comprehensive Approach Towards Data Preprocessing Techniques and Association Rules. *Bharati Vidyapeeth's Institute of Computer Applications and Management (BVICAM)*, 12.

Sivakumaran, T. a., Igo, R. P., Kidd, J. M., Itsara, A., Kopplin, L. J., Chen, W., Hagstrom, S. a., Peachey, N. S., Francis, P. J., Klein, M. L., Chew, E. Y., Ramprasad, V. L., Tay, W. T., Mitchell, P., Seielstad, M., Stambolian, D. E., Edwards, A. O., Lee, K. E., Leontiev, D. V., … Iyengar, S. K. (2011). A 32 kb critical region excluding Y402H in CFH mediates risk for age-related macular degeneration. *PLoS ONE*, *6*(10). https://doi.org/10.1371/journal.pone.0025598

Skryjomski, P., & Krawczyk, B. (2017). Influence of minority class instance types on SMOTE imbalanced data oversampling. *Proceedings of Machine Learning Research LIDTA 2017*, *74*, 7–21. http://proceedings.mlr.press/v74/skryjomski17a/skryjomski17a.pdf

Sobrin, L., & Seddon, J. M. (2014). Nature and nurture- genes and environment- predict onset and progression of macular degeneration. In *Progress in Retinal and Eye Research* (Vol. 40, pp. 1–15). https://doi.org/10.1016/j.preteyeres.2013.12.004

Soft computing and intelligent information systems (SCI2S). (2020). *Noisy Data in Data Mining - Soft Computing and Intelligent Information Systems*. Sci2S.Ugr.Es. https://sci2s.ugr.es/noisydata

Sotoca, V., García, V., Sánchez, J., & Mollineda, R. (2007). The class imbalance problem in pattern classification and learning. *Data Engineering*, *VIII*, 352.

http://marmota.dlsi.uji.es/WebBIB/papers/2007/1_GarciaTamida2007.pdf

Spencer, K. L., Olson, L. M., Schnetz-Boutaud, N., Gallins, P., Agarwal, A., Iannaccone, A., Kritchevsky, S. B., Garcia, M., Nalls, M. A., Newman, A. B., Scott, W. K., Pericak-Vance, M. A., & Haines, J. L. (2011). Using genetic variation and environmental risk factor data to identify individuals at high risk for age-related macular degeneration. *PLoS ONE*, *6*(3), e17784. https://doi.org/10.1371/journal.pone.0017784

Stone, E. M., Aldave, A. J., Drack, A. V., MacCumber, M. W., Sheffield, V. C., Traboulsi, E., & Weleber, R. G. (2012). Recommendations for genetic testing of inherited eye diseases: Report of the American academy of ophthalmology task force on genetic testing. *Ophthalmology*, *119*(11), 2408–2410. https://doi.org/10.1016/j.ophtha.2012.05.047

Sturm, M., Gehrke, J., Elhadad, N., Lou, Y., Caruana, R., & Koch, P. (2015). *Intelligible Models for HealthCare*. 1721–1730. https://doi.org/10.1145/2783258.2788613

Sun, Z., Song, Q., Zhu, X., Sun, H., Xu, B., & Zhou, Y. (2015). A novel ensemble method for classifying imbalanced data. *Pattern Recognition*, *48*(5), 1623–1637. https://doi.org/10.1016/j.patcog.2014.11.014

Tan, P.-N., Steinbach, M., Karpatne, A., & Kumar, V. (2018). *Introduction to Data Mining (2nd Edition)* (2nd ed.). Pearson.

Tejera, E., Jose Areias, M., Rodrigues, A., Rama, A., Manuel Nieto-Villar, J., & Rebelo, I. (2011). Artificial neural network for normal, hypertensive, and preeclamptic pregnancy classification using maternal heart rate variability indexes. *Journal of Maternal-Fetal and Neonatal Medicine*, *24*(9), 1147–1151. https://doi.org/10.3109/14767058.2010.545916

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: A retrospective. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *73*(3), 273–282. https://doi.org/10.1111/j.1467-9868.2011.00771.x

Varshney, K. R., Khanduri, P., Sharma, P., Zhang, S., & Varshney, P. K. (2018). Why Interpretability in Machine Learning? An Answer Using Distributed Detection and Data Fusion Theory. *2018 ICML Workshop on*

Human Interpretability in Machine Learning. https://arxiv.org/pdf/1806.09710.pdf

Velikova, M., Van Scheltinga, J. T., Lucas, P. J. F., & Spaanderman, M. (2014). Exploiting causal functional relationships in Bayesian network modelling for personalised healthcare. *International Journal of Approximate Reasoning*, *55*(1 PART 1), 59–73. https://doi.org/10.1016/j.ijar.2013.03.016

Villa, P. M., Marttinen, P., Gillberg, J., Inkeri Lokki, A., Majander, K., Ordén, M. R., Taipale, P., Pesonen, A., Räikkönen, K., Hämäläinen, E., Kajantie, E., & Laivuori, H. (2017). Cluster analysis to estimate the risk of preeclampsia in the high-risk Prediction and Prevention of Preeclampsia and Intrauterine Growth Restriction (PREDO) study. *PLoS ONE*, *12*(3), 1–14. https://doi.org/10.1371/journal.pone.0174399

Vluymans, S. (2019). Learning from imbalanced data. In *Studies in Computational Intelligence* (Vol. 807, pp. 81–110). Springer Verlag. https://doi.org/10.1007/978-3-030-04663-7_4

Weiss, G. M., & Provost, F. (2003). Learning When Training Data are Costly: The Effect of Class Distribution on Tree Induction. *Journal of Artificial Intelligence Research*, *19*, 315–354. https://doi.org/10.1613/jair.1199

Weiss, G., McCarthy, K., & Zabar, B. (2007). Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs? *Dmin*, 1–7. http://storm.cis.fordham.edu/~gweiss/papers/dmin07-weiss.pdf

Wong, W. L., Su, X., Li, X., Cheung, C. M. G., Klein, R., Cheng, C.-Y., & Wong, T. Y. (2014). Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis. *The Lancet. Global Health*, *2*(2), e106-16. https://doi.org/10.1016/S2214-109X(13)70145-1

World Medical Associationdeclaration of Helsinki. (2014). World medical association declaration of Helsinki: Ethical principles for medical research involving human subjects. In *Journal of the Korean Medical Association* (Vol. 57, Issue 11). https://doi.org/10.5124/jkma.2014.57.11.899

Xiao, C., Choi, E., & Sun, J. (2018). Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, *25*(10), 1419–1428.

Yang, X., Hu, J., Zhang, J., & Guan, H. (2010). Polymorphisms in CFH, HTRA1 and CX3CR1 confer risk to exudative age-related macular degeneration in Han Chinese. *Br J Ophthalmol*, *94*(9), 1211–1214. https://doi.org/10.1136/bjo.2009.165811

Zayyad, M. A., & Toycan, M. (2018). Factors affecting sustainable adoption of e-health technology in developing countries: an exploratory survey of Nigerian hospitals from the perspective of healthcare professionals. *PeerJ*, *6*, e4436. https://doi.org/10.7717/peerj.4436

Zhang, M., & Baird, P. N. (2016). A decade of age-related macular degeneration risk models: What have we learned from them and where are we going? *Ophthalmic Genetics*, *00*(00), 1–7. https://doi.org/10.1080/13816810.2016.1227451

Zhou, S.-M., & Gan, J. Q. (2008). Low-level interpretability and high-level interpretability: a unified view of data-driven interpretable fuzzy system modelling. *Fuzzy Sets and Systems*, *159*(23), 3091–3131. https://doi.org/10.1016/j.fss.2008.05.016

Zhou, Z. H. (2012). Ensemble methods: Foundations and algorithms. In Chapman and Hall/CRC (Ed.), *Ensemble Methods: Foundations and Algorithms*. CRC Press. https://doi.org/10.1201/b12207

Zhu, B., Gao, Z., Zhao, J., & vanden Broucke, S. K. L. M. (2019). IRIC: An R library for binary imbalanced classification. *SoftwareX*, *10*(October), 100341. https://doi.org/10.1016/j.softx.2019.100341