Desarrollo de Modelos de Aprendizaje Automático para Detección de Apnea del Sueño

Elvis Yael De los Santos Lopez [1]

INTRODUCCIÓN

El sueño es un componente esencial para el desarrollo de estilos de vida saludable; sin embargo, a nivel mundial, al menos 45% de la población padece algún tipo de trastorno del sueño. Los Trastornos Respiratorios del Sueño (TRS) son los trastornos del sueño que se presentan con mayor frecuencia (American Academy of Sleep Medicine, 2024; Álvarez García & Jiménez Correa, 2020).

El TRS más común es la Apnea Obstructiva del Sueño (AOS) que se caracteriza por episodios recurrentes de cese o disminución de la circulación del aire durante el sueño ocasionando una disminución en el aporte de oxígeno que desencadena múltiples consecuencias a nivel físico y metabólico y afecta la calidad de vida de las personas que la padecen (Yathish & Manjula, 2024).

Existe evidencia científica que reporta relación entre los TRS y enfermedad cardiovascular, así como diversos desórdenes metabólicos entre los que destacan resistencia a la insulina, hipertensión arterial, diabetes y obesidad (Ott et al., 2017; Yathish & Manjula, 2024).

A pesar de su alta prevalencia e impacto en la salud pública, los Trastornos Respiratorios del Sueño (TRS), y en particular la Apnea Obstructiva del Sueño (AOS), continúan siendo sub diagnosticados y, en muchos casos, tratados de forma tardía. Esto se debe a múltiples factores, entre los que destacan el acceso limitado a estudios especializados como la polisomnografía. La ausencia de un diagnóstico oportuno conlleva a la progresión silenciosa de complicaciones cardiovasculares y metabólicas, incrementando la carga económica y social tanto para los sistemas de salud como para los pacientes.

Esta situación resalta la necesidad urgente de estrategias que mejoren la detección temprana de la AOS, especialmente en poblaciones con difícil acceso a servicios médicos especializados.

MATERIALES Y METODOLOGÍA

La metodología aplicada adopta un enfoque estructurado de aprendizaje automático para la construcción de un modelo predictivo en un entorno clínico (Shamout et al., 2021).

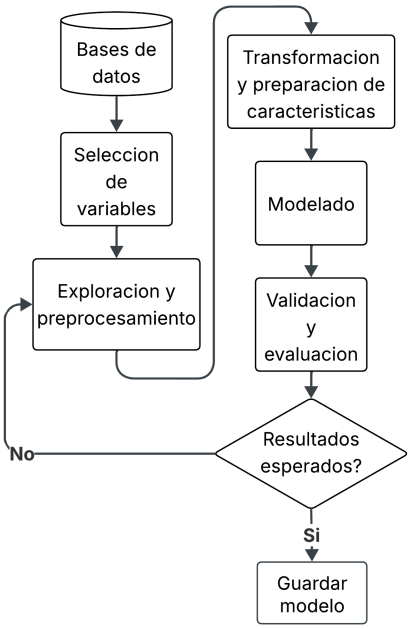

Figura 1. Metodologia.

Materiales y Fuentes de Datos

Los conjuntos de datos utilizados en este estudio fueron solicitados a través del

National Sleep Research Resource (NSRR), un repositorio especializado en la investigación del sueño patrocinado por el National Heart, Lung, and Blood Institute (Zhang et al., 2018). Se emplearon dos bases de datos complementarias para el entrenamiento y validación del modelo:

1. **Sleep Heart Health Study (SHHS)**

   Diseñado para explorar las consecuencias cardiovasculares de los trastornos respiratorios del sueño, este estudio longitudinal y multicéntrico recopiló datos clínicos y polisomnográficos de 6,441 participantes mayores de 40 años reclutados entre 1995-1998 (Quan et al., 1997). El SHHS representa una fuente de datos con diversidad étnica, aunque con predominio de población caucásica.

2. **Hispanic Community Health Study/Study of Latinos (HCHS/SOL)**

   Estudio multicéntrico que incluyó a 16,000 adultos hispanos/latinos de 18-74 años entre 2008-2011, de los cuales aproximadamente 14,000 completaron estudios de sueño domiciliarios (Redline et al., 2014). Este dataset proporciona datos representativos de hispanos/latinos en Estados Unidos, permitiendo ampliar la diversidad étnica del modelo.

Selección de variables

Las variables seleccionadas para el desarrollo del modelo corresponden a los factores de riesgo para AOS y su asociación con la severidad de la misma:

1. **Variables demográficas**

   La edad y el sexo son factores clave en la epidemiología de la Apnea Obstructiva del Sueño (AOS): es más común en hombres y su prevalencia aumenta con la edad. (Peppard et al., 2013).

2. **Variables antropométricas**

   índice de masa corporal (IMC), medida con fuerte asociación con la prevalencia y severidad de AOS. (Franklin & Lindberg, 2015).

3. **Variables cardiovasculares**

   Presión arterial sistólica y diastólica, relacionadas con la elevada comorbilidad entre hipertensión y AOS. (Javaheri et al., 2017).

4. **Factores de riesgo adicionales**

   El tabaquismo puede empeorar la inflamación de la vía aérea superior, lo que contribuye a la fisiopatología de la Apnea Obstructiva del Sueño (AOS). (Mayo Clinic, 2023).

5. **Variable objetivo**

   El Índice de Apnea-Hipopnea (IAH) representa el promedio de apneas e hipopneas por hora de sueño. La variable específica 'nsrr_ahi_hp3r_aasm15' sigue los criterios de la AASM de 2015, considerados el estándar actual para el diagnóstico y la clasificación de la severidad de la AOS. (Kingshott, 2017).

Exploración y Preprocesamiento de Datos

El análisis exploratorio de datos representa un paso crítico para comprender la distribución, estructura y particularidades de las bases de datos. (Konopka et al., 2018)

- **Combinación de datasets**

Se integraron las bases SHHS y HCHS mediante homologación de variables.

- **Ajuste de la Distribución de Datos**

  Se implementó un muestreo estratificado por edad y sexo para equilibrar la representación de diferentes niveles de severidad de AOS, siguiendo recomendaciones metodológicas para el manejo de clases desbalanceadas en tareas de clasificación dentro del ámbito médico. (He & Garcia, 2009)

- **Análisis de valores faltantes**

  La eliminación de filas con valores faltantes en variables fundamentales e imputación de los datos numéricos mediante K-Nearest Neighbors (KNN) (Konopka et al., 2018).

2. Transformación y Preparación de Características

La transformación de datos clínicos requiere considerar la naturaleza específica de cada variable y su interpretabilidad clínica (Shickel et al., 2017):

- **Manejo de valores atípicos**

  Se adoptó un enfoque robusto basado en el rango intercuartílico (IQR), utilizando un factor ampliado de 3×IQR en lugar del umbral convencional de 1.5×IQR. Esta decisión metodológica se fundamenta en la naturaleza de los datos médicos, donde los valores extremos pueden reflejar condiciones clínicas legítimas y no necesariamente errores o ruido estadístico (Cousineau & Chartier, 2010).

- **Creación de características derivadas**

  Se generaron variables clínicamente relevantes siguiendo estándares médicos internacionales, como la categorización del IMC según la OMS (Organización Mundial de la Salud, 2024) y la clasificación de la presión arterial basada en guías de la American Heart Association (Heart Association, 2024).

- **Ingeniería de características**

  Se agregaron variables binarias como edad_obesidad_risk (para pacientes mayores de 50 años con IMC ≥ 30) y hombre_edad_media (hombres entre 40 y 65 años) (Peppard et al., 2013). Adicionalmente se agrega la variable clinical_risk_score como un puntaje combinado que suma la presencia de factores de riesgo clave como obesidad, hipertensión y tabaquismo (Franklin & Lindberg, 2015). Se subraya la importancia de la ingeniería de características en la mejora de la interpretabilidad y el rendimiento de los modelos predictivos (Zheng & Casari, 2018).

- **Selección de variable predictora para el entrenamiento**

  Selección de la variable objetivo o variable predictora y se excluyen variables redundantes. La variable objetivo utilizada en el entrenamiento binario es ‘apnea’, donde 1 indica la presencia de apnea y 0 pacientes sanos. Para el entrenamiento multiclase se usa la variable 'apnea_severity_ordinal', que codifica numéricamente los cuatro niveles de severidad.

- **Codificación de variables categóricas**

  Se aplicaron técnicas de codificación numérica que preservan la

interpretabilidad clínica de las variables.

- **Estandarización**

  La estandarización es una práctica estándar en el aprendizaje automático, garantiza que no existan diferentes escalas en las variables numéricas, lo que podría afectar negativamente a modelos sensibles a escalas como SVM (Shamout et al., 2021).

- **Balanceo de clases**

  Para mitigar el desbalance en las distribuciones de clases del dataset, se implementaron estrategias de balanceo. Este paso es crucial, ya que los modelos de Machine Learning son sensibles a las distribuciones desbalanceadas, lo que puede comprometer su rendimiento (He & Garcia, 2009).

3. Modelado Predictivo

La selección de algoritmos se basó en su rendimiento documentado en problemas similares de clasificación en medicina del sueño (Park et al., 2025):

- **Random Forest**

  Random Forest, ha sido ampliamente utilizado en el área de la salud y ha demostrado ser eficiente en tareas de clasificación y predicción (Breiman, 2001). Su capacidad para combinar múltiples árboles de decisión reduce el sobreajuste, gestiona el ruido de los datos y captura relaciones complejas entre variables, características cruciales dada la variabilidad inherente en los datos clínicos.

- **Gradient Boosting**

  Esta técnica de ensamble, ha demostrado un rendimiento notable en tareas de clasificación, incluso cuando se trabaja con datos limitados (Friedman, 2001). Su aproximación iterativa para construir modelos secuencialmente le permite corregir los errores de los modelos previos, generando predicciones para el análisis en el área de la salud.

- **Support Vector Machine (SVM)**

  El algoritmo Support Vector Machine (SVM) se destaca por su desempeño en diversas aplicaciones en el área de la salud (Park et al., 2025). Es particularmente efectivo en tareas de clasificación, ya que busca el hiperplano que no solo divide los datos, sino que lo hace con el mayor margen de separación posible entre las clases. (Sidey-Gibbons & Sidey-Gibbons, 2019).

4. Validación y Evaluación

La validación es fundamental en modelos destinados a aplicaciones médicas:

- **Validación cruzada estratificada**

  Se implementó un esquema de 5 particiones manteniendo la proporción de clases en cada fold (pliegue) (Steyerberg & Harrell, 2016).

- **Métricas orientadas a aplicación clínica**

  Uso de métricas centradas en la capacidad de detección, como la sensibilidad recall y el área bajo la curva AUC-ROC (Fawcett, 2006). En el contexto de la AOS, las consecuencias asociadas a los falsos negativos, es decir, no identificar a un paciente con AOS, son clínicamente más relevantes

que los falsos positivos (Steyerberg & Harrell, 2016).

- **Optimización de hiper parámetros**

  Se empleó RandomizedSearchCV en el espacio de parámetros definido mediante rangos, con énfasis en maximizar el recall y AUC-ROC (Fawcett, 2006).. RandomizedSearchCV es una tecnica ampliamente utilizada para la búsqueda de parámetros, se define un rango de parámetros, el número de pliegues y la cantidad de iteraciones, donde el algoritmo se ejecuta sobre este espacio de parámetros para encontrar las mejores combinaciones de parámetros (Arindam, 2022b).

PROPUESTA

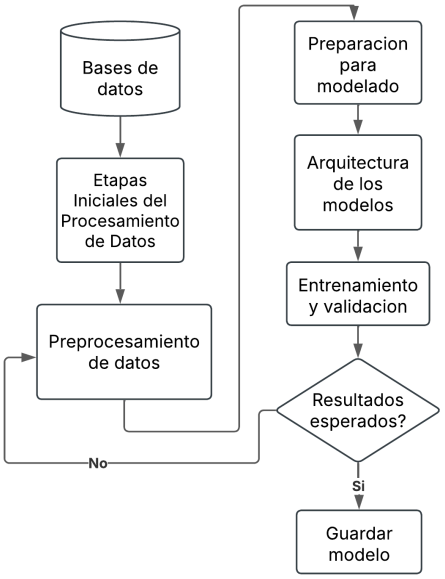

Figura 2. Pipeline para el Diagnóstico de AOS

Conectar una flecha de el entrenamiento y validación al preprocesamiento

1. Etapas Iniciales del Procesamiento de Datos

Se describen las etapas previas al preprocesamiento, fase donde se estudian los datos y su distribución para elegir las técnicas que se aplicaran en el preprocesamiento.

1.1 Integración de bases de datos

La combinación de las bases de datos Sleep Heart Health Study (SHHS) y Hispanic Community Health Study/Study of Latinos (HCHS/SOL) maximiza la diversidad étnica y el tamaño muestral del modelo predictivo, ampliando la representatividad poblacional al incluir múltiples razas. Ambas bases de datos comparten variables equivalentes con nomenclaturas diferentes, lo que permite su integración efectiva.

El proceso de integración incluye lo siguiente:

1. Carga simultánea de ambas bases de datos.
2. Homologación de variables mediante mapeo de nomenclaturas.
3. Eliminacion de registros con valores faltantes en variables clave: Índice de apnea-hipopnea, Índice de masa corporal, sexo y edad. Un paso fundamental en el preprocesamiento de datos, ya que estas variables representan características fundamentales para el entrenamiento del modelo.

   La eliminación de estos registros resultó en una reducción significativa del tamaño de la base de datos. Sin embargo, se justifica dado que los registros eliminados no pueden imputarse de manera confiable, y su eliminación garantiza la integridad de los datos utilizados para el entrenamiento del modelo.

4. Almacenamiento en formato CSV para su posterior procesamiento.

1.2 Análisis exploratorio de datos

El análisis exploratorio de datos (EDA), una fase crucial en la metodología de nuestro estudio, reveló la presencia de patrones esperados dentro del conjunto de datos.

- **Valores Faltantes**

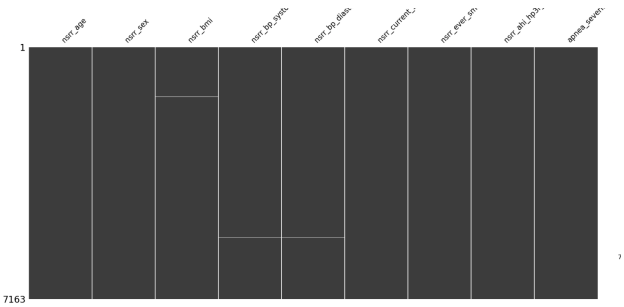

Figura 3. Valores faltantes en el dataset antes del preprocesamiento, representados por las líneas blancas horizontales.

Dado que los valores faltantes representan menos del 5% del conjunto total de datos, se optó por su imputación utilizando el método KNN (K-Nearest Neighbors), considerado apropiado para proporciones bajas de datos faltantes (Park et al., 2025).

- **Distribución de variables numéricas:**

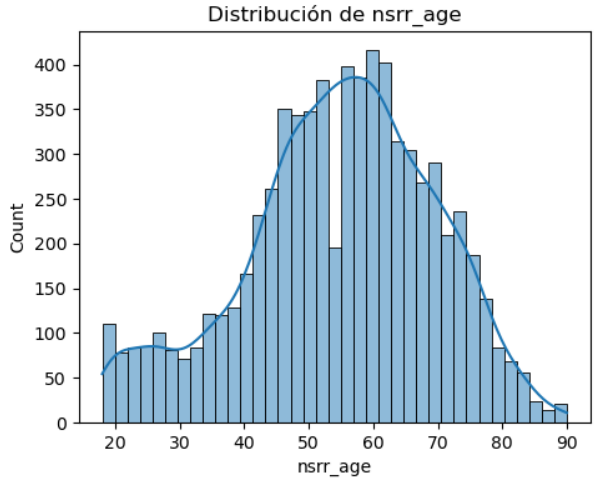

Figura 4. Edad de los pacientes: distribución antes del preprocesamiento.

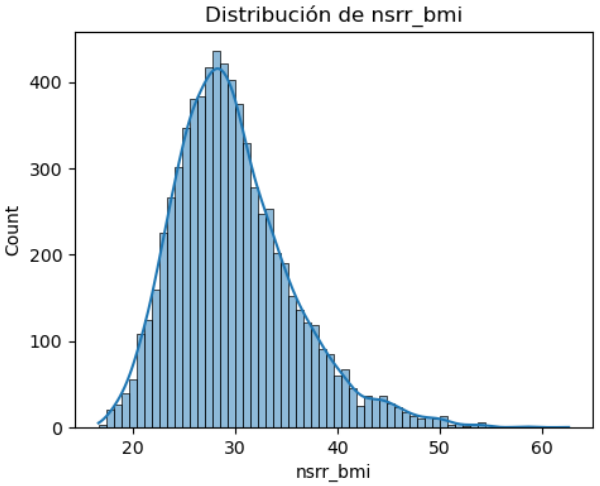

Figura 4.1. Índice de masa corporal (IMC): distribución antes del preprocesamiento.

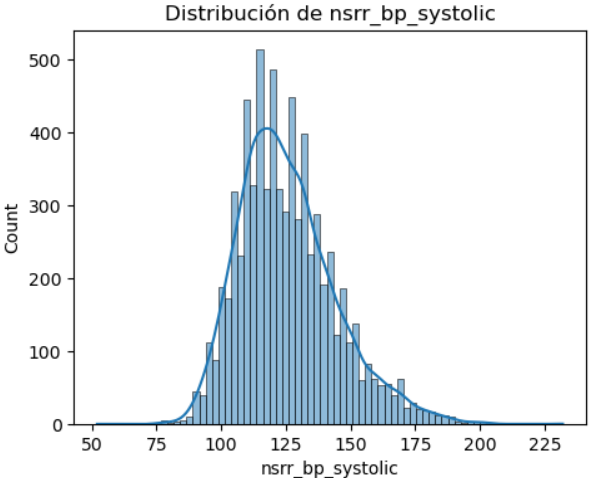

Figura 4.2. Presión sistólica: distribución antes del preprocesamiento.

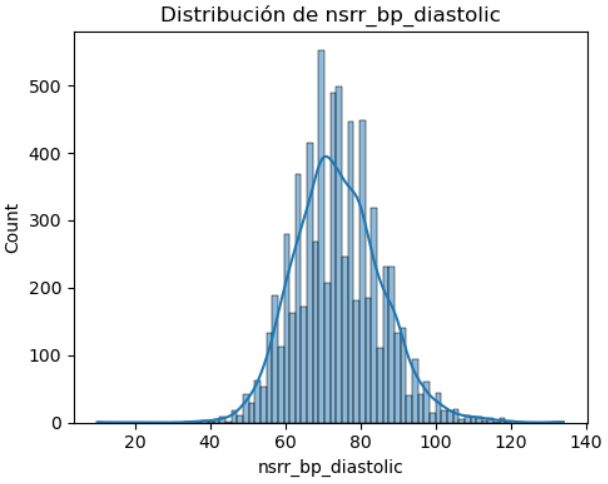

Figura 4.3. Presión diastólica: distribución antes del preprocesamiento.

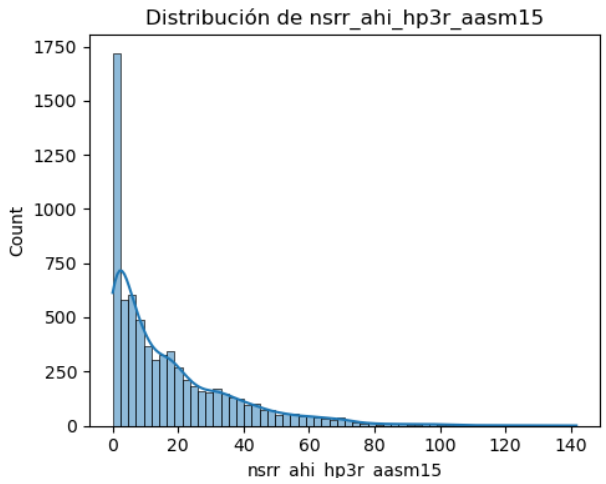

Figura 4.4. Índice de apnea-hipopnea: distribución antes del preprocesamiento.

Las gráficas muestran un sesgo en el Índice de Masa Corporal y el Índice de Apnea-Hipopnea, patrón esperado dada la asociación entre obesidad y severidad de AOS documentada en la literatura (Franklin & Lindberg, 2015. Las variables demográficas, particularmente la edad, y las variables cardiovasculares como la presión arterial sistólica y diastólica, presentan distribuciones con ligero sesgo, indicando un conjunto de datos balanceado que requiere intervenciones de preprocesamiento mínimas.

1.3 Ajuste de la Distribución de Datos

En esta fase se balancea la distribución de clases del conjunto de datos antes del entrenamiento del modelo. Las clases a predecir: pacientes sin apnea, con apnea leve, moderada y severa, se definen a partir del índice nsrr_ahi_hp3r_aasm15, el cual aplica los criterios establecidos por la AASM en su revisión de 2015 (Heart Association, 2024). Este índice clasifica a los pacientes como: normales (IAH < 5 eventos por hora), con AOS leve (IAH entre 5 y 14.9), moderada (IAH entre 15 y 29.9) y severa (IAH ≥ 30).

El proceso completo consiste en los siguientes pasos:

1. Carga de los datasets SHHS y HCHS.

2. Clasificación inicial de los pacientes con base en el índice de apnea-hipopnea, lo que permite agrupar los casos según los cuatro niveles de severidad definidos.

3. Análisis de la distribución original de clases, que permite visualizar y cuantificar el desbalance presente en los datos sin procesar.

4. Ajuste de la distribución mediante muestreo estratificado, utilizando las variables de edad y sexo como criterios de estratificación. Esta estrategia no solo mantiene la representatividad demográfica, sino que también mejora la capacidad del modelo para generalizar, al asegurar una cobertura adecuada de cada subgrupo poblacional (Patiño-Pérez et al., 2023).

   La distribución objetivo de clases se define dentro de los siguientes rangos:

   - Pacientes sin apnea: 28% – 35%
   - Apnea leve: 25% – 30%
   - Apnea moderada: 21% – 25%
   - Apnea severa: 21% – 25%.

5. Almacenamiento del dataset procesado con la nueva distribución balanceada para su uso en el script principal de entrenamiento.

Este procesamiento adicional es esencial para garantizar que el modelo reciba un conjunto de datos representativo y balanceado. La estratificación refuerza la integridad del análisis clínico al considerar factores relevantes (Patiño-Pérez et al., 2023).

2. Preprocesamiento de datos

La fase de preprocesamiento corresponde a la segunda etapa de la investigación, priorizando la preservación de información clínica relevante y la calidad del dataset para el entrenamiento. Se adoptó un enfoque que minimiza la introducción de datos no relacionados a las bases de datos (Rajkomar et al., 2019). Las técnicas seleccionadas se eligieron con base en su impacto sobre los datos, favoreciendo aquellas que permiten su uso sin comprometer la validez clínica ni la distribución original de las variables.

2.1 Conversión de variables categóricas

La conversión de variables categóricas a un formato numérico fue necesaria para garantizar la compatibilidad con los algoritmos de aprendizaje automático. Se empleó codificación ordinal para variables binarias (como sexo y tabaquismo) y tratamiento explícito de valores como "Not reported", que se transformaron en NaN para ser imputados posteriormente usando KNN.

2.2 Gestion de valores faltantes

Debido a la cantidad mínima de valores faltantes, se optó por usar KNN en lugar de la eliminación directa. La imputación con K-Nearest Neighbors (KNN) considera las relaciones entre variables, al imputar datos en función de pacientes similares. Esta elección permite mantener la consistencia interna del dataset mientras se reduce la pérdida de información (Nagarajan & Dhinesh Babu, 2022).

2.3 Detección y tratamiento de outliers (valores atípicos)

Dada la naturaleza de los datos médicos y la condición que se busca predecir, el tratamiento de valores atípicos requirió un enfoque más permisivo. Para la detección de los outliers se utilizó un rango intercuartílico ampliado (IQR × 3) en lugar del estándar 1.5 para identificar valores extremos que podrían representar condiciones clínicas reales (Rajkomar et al., 2019).

2.4 Filtrado por edad

Se filtraron los registros de pacientes fuera del rango de 30 a 65 años, con base en estudios que identifican este rango como el de mayor prevalencia de AOS (Vensel-Rundo, 2019). Este filtrado busca centrar el entrenamiento en la población más afectada por la AOS.

2.5 Creación de características

En esta sección se crean variables que aportan información útil al modelo basándonos en información existente y convenciones de organizaciones reconocidas como la World Health Organization (Organización Mundial de la Salud) y la American Academy of Sleep Medicine (AASM).

2.5.1 Categorización del IMC

La categorización del IMC, basada en los estándares de la Organización Mundial de la Salud (OMS, 2024), proporciona una interpretación clínica adicional que complementa y enriquece el valor numérico original.

1. IMC <= 18.5: Bajo
2. IMC >18.5 y <=25: Normal
3. IMC >25 y <=30: Sobrepeso
4. IMC >30 y <=35: Obesidad I
5. IMC >35 y <=40: Obesidad II
6. IMC >40: Obesidad III

2.5.2 Categorización de la presión arterial

La clasificación de presión arterial se basa en la guía de la American Heart Association (Heart Association, 2024):

1. PA Sistólica < 120 y PA Diastólica < 80: Normal

2. PA Sistólica >= 120 y PA Sistólica <= 129 o PA Diastólica < 80: Elevada
3. PA Sistólica >= 130 y PA Sistólica <= 139 o PA Diastólica >= 80 y PA Diastólica <= 89: Hipertensión Etapa 1
4. PA Sistólica >= 140 y PA Sistólica <= 180 o PA Diastólica >= 90 y PA Diastólica <= 120: Hipertensión Etapa 2
5. PA Sistólica > 180 o PA Diastólica > 120: Crisis Hipertensiva

La función considera tanto la presión sistólica como diastólica. Esta variable es especialmente relevante debido a que la hipertensión tiene una relación bidireccional con la AOS (Chávez-González & Soto, 2018).

2.6 Creación de Relaciones Mediante Ingeniería de Características

Se utilizan técnicas de ingeniería de características para generar nuevas variables que pudieran capturar relaciones significativas (Zheng & Casari, 2018).

2.6.1 Relación entre edad y IMC

Variable diseñada para señalar a aquellos individuos que presentan un riesgo elevado en base a su edad y su índice de masa corporal. Se asigna un valor de 1 si el paciente tiene más de 50 años y su Índice de Masa Corporal (IMC) es más de 30. La combinación de una edad avanzada y la obesidad es un factor de riesgo reconocido en diversas condiciones de salud (Franklin & Lindberg, 2015).

2.6.2 Relacion entre Sexo y edad

Para capturar la interacción específica entre el sexo masculino y el riesgo cardiovascular en la mediana edad, creamos una variable binaria denominada hombre_edad_media. Esta variable funciona como un indicador de riesgo dirigido que identifica a los hombres que se encuentran en el rango de edad de mayor vulnerabilidad cardiovascular (Franklin & Lindberg, 2015).

2.6.3 Factores de riesgo combinados

Debido a que los factores de riesgo cardiovascular no actúan de forma aislada, desarrollamos una métrica llamada clinical_risk_score que cuantifica el riesgo presente en cada paciente.

El cálculo del puntaje se realiza mediante la suma de tres factores de riesgo presentes en el dataset: la obesidad (determinada por el índice de masa corporal), la hipertensión, y el tabaquismo activo (identificado a través de la variable nsrr_current_smoker). Cada uno de estos componentes aporta una unidad al puntaje.

Valores más elevados indican una mayor acumulación de factores de riesgo conocidos, como lo han demostrado estudios previos en el campo de la medicina (Franklin & Lindberg, 2015).

2.7 Creación de variables objetivo

La categorización de la severidad se realizó con base en los criterios establecidos por la American Academy of Sleep Medicine (AASM) para el Índice de Apnea-Hipopnea (IAH) (American Academy of Sleep Medicine, 2012). Los puntos de corte utilizados fueron los siguientes:

- nsrr_ahi_hp3r_aasm15 < 5: Normal
- 5 >= nsrr_ahi_hp3r_aasm15 < 15: Leve
- 15 >= nsrr_ahi_hp3r_aasm15 < 30: Moderada
- nsrr_ahi_hp3r_aasm15 >= 30: Severa

Los puntos de corte (5, 15, 30) fueron establecidos por la American Academy of Sleep Medicine (AASM) para clasificar la severidad de la apnea del sueño.

1. **Dos variables:** Se crea una variable categórica que indica la severidad, útil para entrenamiento multiclase, así como una variable binaria

(presencia/ausencia) que se utiliza para la clasificación binaria.

2.8 Resultados del preprocesamiento

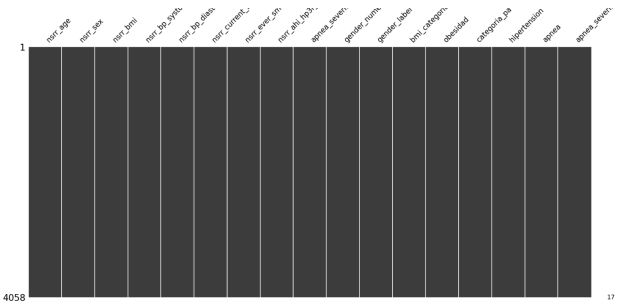

Figura 5. Valores faltantes en el dataset después del preprocesamiento

El preprocesamiento de los datos resultó en una reducción del 14.5% del dataset, pasando de 7,163 a 6,121 registros. Principalmente por la eliminación de valores atípicos y la aplicación de un filtro de edad. Tras este proceso, el conjunto de datos final quedó libre de valores faltantes, lo que se detalla visualmente en la Figura 5.

- **Distribución de variables numéricas**

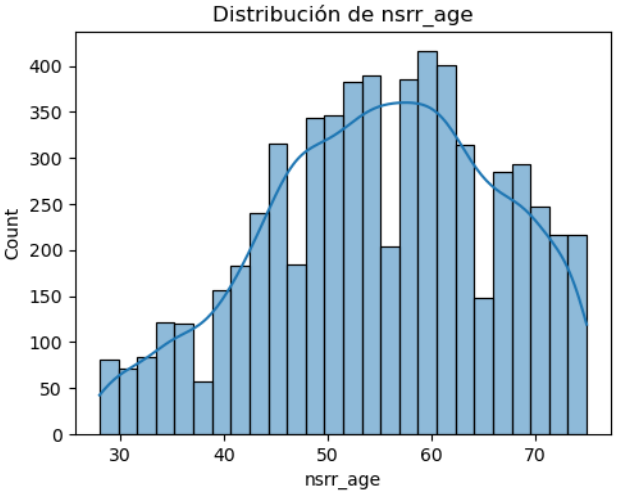

Figura 6. Edad de los pacientes: distribución posterior al preprocesamiento.

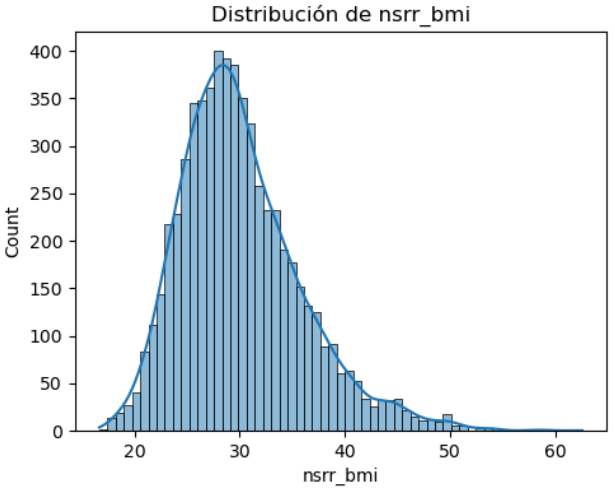

Figura 6.1. Índice de masa corporal (IMC): distribución posterior al preprocesamiento

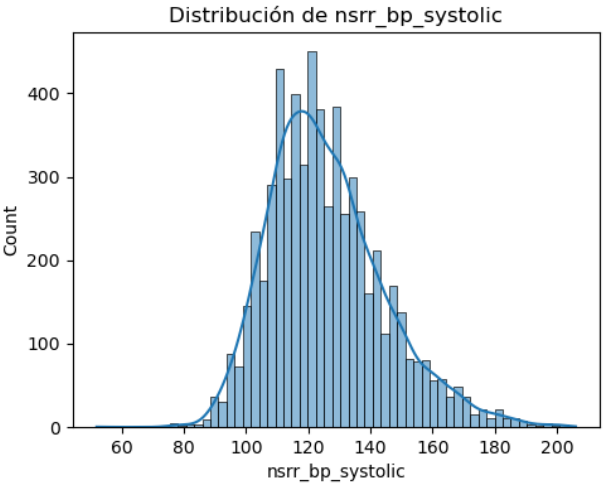

Figura 6.2. Presión sistólica: distribución posterior al preprocesamiento.

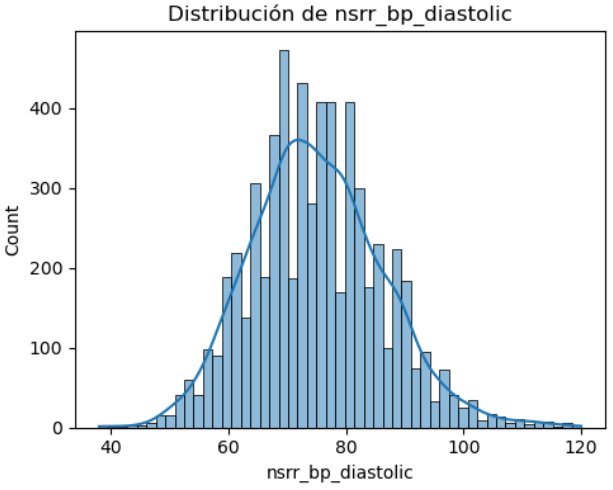

Figura 6.3. Presión diastólica: distribución posterior al preprocesamiento.

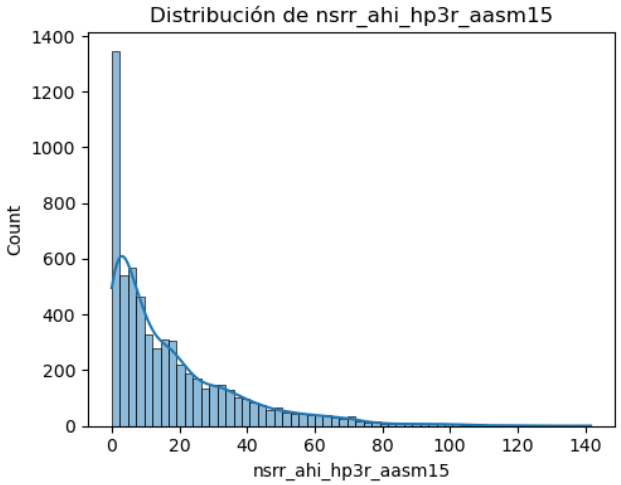

Figura 6.4. Índice de apnea-hipopnea (IAH): distribución posterior al preprocesamiento.

Las figuras ilustran el resultado del preprocesamiento en la distribución de variables numéricas. Se observan:

- Sesgos mínimos en variables como edad, coherente con el filtro de edad aplicado.
- Sesgos en IMC y el índice hp3r, un resultado esperado debido al equilibrio de clases y la prevalencia de pacientes con sobrepeso y AOS.

3. Preparación de los datos para modelado

La fase de preparación transforma el conjunto preprocesado en una representación optimizada para el aprendizaje automático, implementando la tercera fase. Esta etapa resulta crucial para garantizar que los modelos puedan extraer patrones relevantes y generalizar adecuadamente a nuevos datos.

3.1 Selección de características

Se selecciona la variable objetivo o variable predictora y se excluyen variables redundantes. La variable objetivo utilizada en el entrenamiento binario es **'apnea'**, donde 1 indica la presencia de apnea y 0 pacientes sanos.

Se crea la variable **'apnea_severity_ordinal'**, que codifica numéricamente los cuatro niveles de severidad: 0 para pacientes sin apnea, 1 para apnea leve, 2 para apnea moderada y 3 para apnea severa. Esto resulta útil para la experimentación, dada la complejidad del diagnóstico de AOS, la clasificación multiclase es un problema complejo que justifica una investigación propia.

Se excluyeron variables redundantes y relacionadas con las variables objetivo para evitar una fuga de información:

- '**nsrr_ahi_hp3u_aasm1**5': Variable que indica el índice de apnea hypopnea, se excluye para evitar filtrado de información al conjunto de entrenamiento.
- '**apnea_severity**': Corresponde a la versión categórica de la variable objetivo por lo que se considera redundante incluirla.
- '**apnea_severity_ordinal**': Se excluye ya que corresponde a la variable objetivo en clasificación multiclase
- '**apnea**': Se excluye ya que corresponde a la variable objetivo en clasificación binaria

3.2 Codificación de variables categóricas

La transformación mediante variables dummy convierte características categóricas en representaciones binarias múltiples, permitiendo que los algoritmos de aprendizaje automático capturen efectivamente las relaciones no lineales entre categorías y el resultado clínico. Este proceso se aplica específicamente a las categorías de IMC y las categorías de presión arterial.

La codificación dummy preserva la información categórica completa mientras facilita la interpretación de los coeficientes del modelo, permitiendo identificar cuáles categorías

contribuyen más a la predicción de severidad de AOS (Zheng & Casari, 2018).

3.3 Normalización de características

La estandarización es una práctica estándar en aprendizaje automático, mediante RobustScaler garantiza que variables con diferentes escalas de medición contribuyen equitativamente al entrenamiento del modelo, implementando.

La normalización transforma cada característica para que presente media 0 y desviación estándar 1, eliminando el sesgo introducido por diferencias de escala (Shamout et al., 2021).

3.4 Balanceo de clases

Para mitigar el desequilibrio de clases presente en el dataset, se aplican técnicas de balanceo específicas únicamente cuando el ratio de desbalance exceda 1.5.

3.4.1 Balanceo binario

Para problemas de clasificación binaria se emplea SMOTE-Tomek, una técnica híbrida que combina sobre muestreo y submuestreo. Este método primero genera muestras sintéticas mediante SMOTE para equilibrar las clases, y posteriormente aplica Tomek Links para identificar y eliminar pares de vecinos cercanos de clases opuestas que pueden generar ambigüedad (Batista et al., 2004).

3.4.2 Balanceo multiclase

Con motivos de experimentación, para el entrenamiento multiclase se usa ADASYN (Adaptive Synthetic Sampling). Esta técnica genera muestras sintéticas priorizando clases minoritarias que resultan más difíciles de clasificar, creando así una distribución más equilibrada y representativa. Como alternativa de respaldo se utiliza SMOTE (Synthetic Minority Over-sampling Technique), que interpola entre ejemplos existentes para crear nuevas instancias sintéticas de las clases subrepresentadas (Batista et al., 2004).

4. Arquitectura de los modelos

Se implementaron tres algoritmos de clasificación, cada uno con características particulares que los hacen potencialmente adecuados para el problema de diagnóstico de AOS. La elección de estos modelos se basa en el desempeño documentado en contextos médicos similares (Park et al., 2025).

4.1 Random Forest

Random Forest constituye un método de ensamble basado en árboles de decisión que construye múltiples árboles utilizando diferentes subconjuntos de datos y características (Mohri et al., 2018). Este algoritmo aborda el sobreajuste típico de árboles individuales mientras mantiene la interpretabilidad relativa de las decisiones del modelo.

Configuración del modelo: El algoritmo se configuró con parámetros que controlan la complejidad del modelo y el manejo del desbalance de clases. Los parámetros clave incluyen el número de estimadores que determina la cantidad de árboles en el ensamble, la profundidad máxima para controlar el sobreajuste, y los criterios de división mínima que regulan la creación de nodos.

4.2 Gradient Boosting

La implementación de Gradient Boosting construye un modelo predictivo mediante el entrenamiento secuencial de modelos débiles, árboles de decisión poco profundos, donde cada iteración se enfoca específicamente en corregir los errores cometidos por iteraciones anteriores. Este enfoque resulta particularmente efectivo para datos con patrones complejos y relaciones no lineales (Mohri et al., 2018).

Configuración del modelo: Los parámetros fundamentales incluyen la tasa de aprendizaje que controla la contribución de cada árbol individual al modelo final, el número de estimadores que define las iteraciones de entrenamiento, y la profundidad máxima de los árboles base para prevenir el sobreajuste. Se configuraron parámetros de submuestreo para introducir regularización adicional y mejorar la generalización del modelo.

4.3 SVM

Support Vector Machine implementa un enfoque fundamentalmente diferente que busca identificar el hiperplano óptimo de separación entre clases en un espacio de características (Mohri et al., 2018). El algoritmo maximiza el margen de separación entre clases, proporcionando robustez ante variaciones en los datos de entrenamiento.

Configuración del modelo: La implementación utiliza un kernel polinomial que permite capturar relaciones no lineales complejas entre las variables predictoras. Los parámetros clave incluyen el parámetro de regularización C que controla el equilibrio entre maximización del margen y minimización del error de clasificación, el parámetro gamma que define la influencia de cada ejemplo de entrenamiento, y el grado del kernel polinomial. Se estableció una tolerancia específica para los criterios de convergencia del algoritmo de optimización.

5. Entrenamiento y Validación

La fase de entrenamiento y validación sigue la metodología propuesta para garantizar una evaluación robusta de los modelos.

5.1 División de datos

La división de datos se hace en un porcentaje de 80% para entrenamiento y el 20% para validación. Esta proporción aporta suficientes datos para el entrenamiento mientras mantiene un conjunto representativo de pruebas para evaluar el rendimiento del modelo en datos no vistos.

5.2 Métricas de evaluación

Se utilizan las siguientes métricas:

- **AUC-ROC (Área Bajo la Curva ROC)**

  El AUC-ROC mide el área bajo esta curva, donde un valor cercano a 1 indica un modelo con excelente capacidad de discriminación o excelente separación entre clases (Fawcett, 2006).

- **Precisión (Precision)**

  La precisión mide la proporción de verdaderos positivos entre todas las predicciones positivas del modelo, cuantificando la confiabilidad de un diagnóstico positivo de AOS (Géron, 2022).

- **Recall (Recuperación/Sensibilidad)**

  En el contexto clínico específico de la AOS, el recall se prioriza sobre la precisión debido a que las consecuencias clínicas de falsos negativos, es decir, no diagnosticar AOS en un paciente que sí la padece, superan las consecuencias de falsos positivos. Los pacientes no diagnosticados permanecen en riesgo de desarrollar complicaciones cardiovasculares y metabólicas severas, mientras que los falsos positivos pueden ser clarificados mediante estudios diagnósticos adicionales (Géron, 2022)

- **F1-score**

  Mide la media armónica de la precisión y el recall, ofreciendo una evaluación integral del rendimiento del modelo que considera tanto la capacidad de

detección como la confiabilidad de las predicciones positivas (Géron, 2022).

5.3 Técnicas de Optimización y Validación

Se describen técnicas de mejora y evaluación más robustas:

1. **Validación cruzada estratificada**

   Para evaluar la robustez y la capacidad de generalización de nuestros modelos, se implementó la validación cruzada StratifiedKFold con 5 particiones (folds).

   Esto es fundamental en datasets con clases desbalanceadas, como es común en el ámbito médico, ya que previene que algunas particiones contengan una cantidad desproporcionada de una clase, lo que puede llevar a estimaciones de rendimiento poco fiables. La validación cruzada estratificada ofrece una evaluación más estable y representativa del desempeño del modelo al garantizar que la distribución de clases se preserve a lo largo de todo el proceso de validación (Géron, 2022).

2. **Búsqueda de hiperparámetros optimizada**

   Se implementa una búsqueda aleatoria (RandomizedSearchCV) sobre el espacio de hiperparámetros con 150 iteraciones por modelo, estrategia que permite explorar configuraciones diversas sin el costo computacional de la búsqueda exhaustiva (Raschka & Mirjalili, 2019).

RESULTADOS

Métricas para modelos con predicciones binarias			
	SVM	**Random Forest**	**Gradient Boost**
Accuracy	0.7943	0.8608	0.8167
Precision	0.7902	0.8776	0.8063
Recall	0.8010	0.8383	0.8333
F1 Score	0.7956	0.8575	0.8196
ROC AUC	0.8702	0.9348	0.8880

1. Support Vector Machine (SVM)

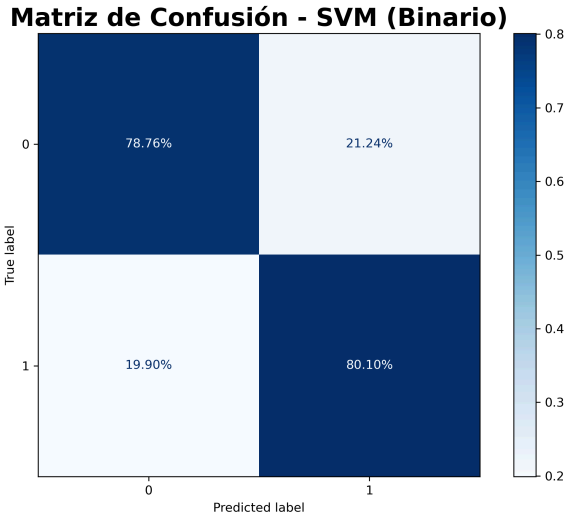

Figura 7. Matriz de confusión del algoritmo SVM con predicciones binarias

El algoritmo SVM alcanzó un rendimiento moderado en la clasificación binaria, con una exactitud (accuracy) del 79.43%. La matriz de confusión para SVM binario muestra un comportamiento equilibrado con 78.76% de verdaderos negativos y 80.10% de verdaderos positivos, aunque presenta 21.24% falsos positivos y 19.9% de falsos negativos. Esta distribución sugiere que el modelo tiene una ligera tendencia a clasificar casos como positivos, lo cual se

refleja en su precisión del 79.02% y sensibilidad (recall) del 80.10%.

El valor F1-score de 0.7956 indica un balance entre precisión y sensibilidad, mientras que el área bajo la curva ROC de 0.8702.

2. Random Forest (RF)

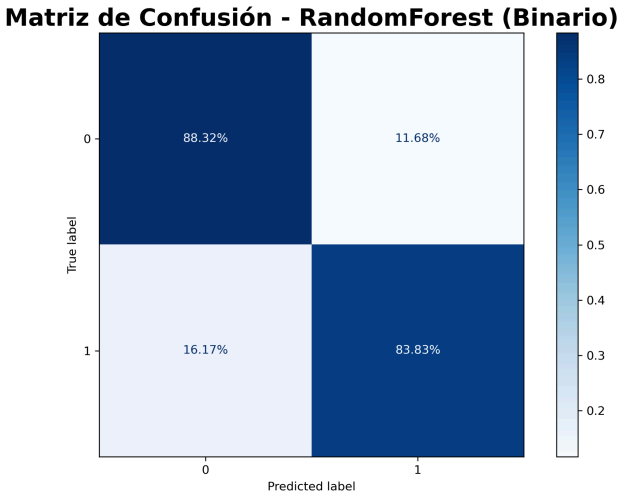

Figura 8. Matriz de confusión del algoritmo Random Forest con predicciones binarias

Random Forest emergió como el algoritmo con mejor rendimiento en clasificación binaria, alcanzando una exactitud del 86.08%. La matriz de confusión correspondiente muestra un 88.32% de verdaderos negativos y 83.83% de verdaderos positivos, con 11.68% de falsos positivos y 16.17% de falsos negativos, precisión del 87.76%, sensibilidad recall del 83.83% y un F1-score de 0.8575.

El área bajo la curva ROC de 0.9348 representa el valor más alto entre los tres algoritmos evaluados, indicando una excelente capacidad de discriminacion entre clases.

3. Gradient Boosting (GB)

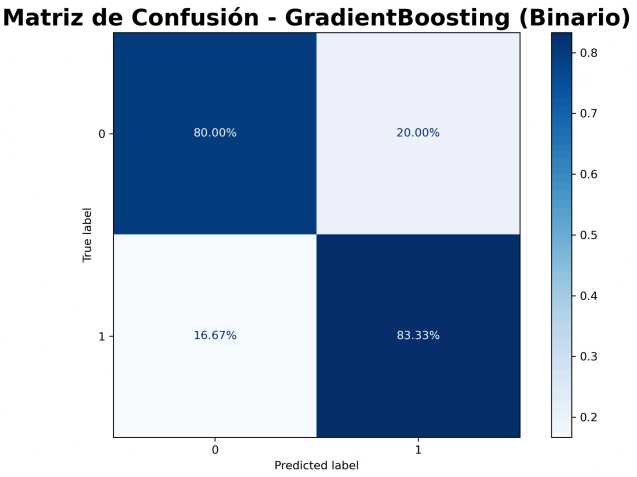

Figura 9. Matriz de confusión del algoritmo Gradient Boost con predicciones binarias

Gradient Boosting mostró un rendimiento intermedio en la clasificación binaria con una exactitud del 81.67%. Su matriz de confusión presenta un 80.0% de verdaderos negativos y 83.33 % de verdaderos positivos, con 20.0% de falsos positivos y 16.67% de falsos negativos. Las métricas resultantes incluyen una precisión del 80.63%, sensibilidad recall del 83.33% y F1-score de 0.8196. Con un área bajo la curva ROC de 0.8880.

DISCUSIÓN

Los resultados obtenidos en este estudio demuestran la viabilidad del uso de algoritmos de aprendizaje automático para un diagnóstico temprano de AOS, siendo Random Forest el algoritmo que mostró el mejor desempeño con una exactitud del 86.08% y un ROC AUC de 0.9348.

La diferencia en el rendimiento entre los algoritmos evaluados refleja las características de cada método. Mientras que SVM mostró las métricas más conservadoras con una exactitud del 79.43%, Gradient Boosting presentó un rendimiento intermedio del 81.67%. Los resultados de Random Forest son prometedores, esto se debe particularmente a su capacidad discriminativa, con el ROC AUC

más alto, lo que indica una mejor separación entre las clases.

Es importante destacar que esta investigación se enfocó en la clasificación binaria debido a las limitaciones inherentes del conjunto de datos utilizado. Se realizaron experimentos de clasificación multiclase para evaluar la capacidad de los algoritmos de distinguir entre los cuatro niveles de severidad de AOS (normal, leve, moderada y severa), sin embargo, los resultados obtenidos fueron considerablemente inferiores a los obtenidos en clasificación binaria: SVM alcanzó una exactitud del 43.48%, Random Forest del 40.58% y Gradient Boosting del 43.72%. Estos valores, todos por debajo del 50%, confirman las limitaciones del dataset para clasificaciones multiclase.

La falta de variables antropométricas utilizadas para la detección de AOS, como la circunferencia del cuello y de la cintura, así como variables fisiológicas adicionales, dificulta la implementación de la clasificación multiclase. La ausencia de estos parámetros en el dataset, limita la capacidad de los algoritmos para realizar clasificaciones sobre la severidad de la AOS.

Los resultados sugieren que, aunque los algoritmos de machine learning pueden ser herramientas valiosas para un diagnóstico inicial de apnea del sueño, la incorporación de variables clínicas adicionales podría mejorar significativamente la precisión diagnóstica y permitir una clasificación más detallada.

CONCLUSIONES

Esta investigación demuestra la viabilidad de usar algoritmos de aprendizaje automático para la detección de AOS. Random Forest se estableció como el método más efectivo, alcanzando una exactitud del 86.08% y un ROC AUC de 0.9348, superando significativamente a SVM (79.43% exactitud) y Gradient Boosting (81.67% exactitud). Estos resultados confirman que Random Forest es más robusto para distinguir entre la presencia y ausencia de apnea del sueño en el conjunto de datos evaluado.

Los hallazgos sugieren que los algoritmos de machine learning representan una herramienta prometedora para el diagnóstico inicial de AOS, especialmente en contextos donde el acceso a estudios polisomnográficos es limitado debido a su elevado costo. Sin embargo, para mejorar el diagnóstico, es fundamental la incorporación de variables clínicas adicionales en futuros desarrollos.

8. Bibliografía

1. Álvarez García, H. B., & Jiménez Correa, U. (2020). Intervención psicológica en trastornos del sueño: una revisión actualizada. *Clínica Contemporánea*, *11*(2). https://doi.org/10.5093/cc2020a9
2. American Academy of Sleep Medicine. (2012). *The AASM manual for the scoring of sleep and associated events: Rules, terminology and technical specifications* (Version 2.0). American Academy of Sleep Medicine.
3. Arindam. (2022, noviembre 2). Hyperparameter tuning using randomized search. *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2022/11/hyperparameter-tuning-using-randomized-search/
4. 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. *SIGKDD Explorations*, *6*(1), 20–29. https://doi.org/10.1145/1007730.1007735
5. Berry, R. B., Budhiraja, R., Gottlieb, D. J., Gozal, D., Iber, C., Kapur, V. K.,

Marcus, C. L., Mehra, R., Parthasarathy, S., Quan, S. F., Redline, S., Strohl, K. P., Ward, S. L. D., & Tangredi, M. M. (2012). Rules for scoring respiratory events in sleep: Update of the 2007 AASM manual for the scoring of sleep and associated events. *Journal of Clinical Sleep Medicine*, *8*(5), 597-619. https://doi.org/10.5664/jcsm.2172

6. Chávez-González, C., & Soto, A. (2018). Evaluación del riesgo de síndrome de apnea obstructiva del sueño y somnolencia diurna utilizando el cuestionario de Berlín y las escalas Sleep Apnea Clinical Score y Epworth en pacientes con ronquido habitual atendidos en la consulta ambulatoria. *Revista Chilena de Enfermedades Respiratorias*, *34*, 19–27.
7. Cousineau, D., & Chartier, S. (2010). Outliers detection and treatment: A review. *International Journal of Psychological Research*, *3*(1), 58-67.
8. Evidently AI. (s. f.). Classification metrics guide. https://www.evidentlyai.com/classification-metrics
9. Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861-874. https://doi.org/10.1016/j.patrec.2005.10.010
10. Franklin, K. A., & Lindberg, E. (2015). Obstructive sleep apnea is a common disorder in the population-a review on the epidemiology of sleep apnea. *Journal of Thoracic Disease*, *7*(8), 1311–1322. https://doi.org/10.3978/j.issn.2072-1439.2015.06.11
11. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263-1284. https://doi.org/10.1109/TKDE.2008.239
12. Javaheri, S., Barbe, F., Campos-Rodriguez, F., Dempsey, J. A., Khayat, R., Javaheri, S., Malhotra, A., Martinez-Garcia, M. A., Mehra, R., Pack, A. I., Polotsky, V. Y., Redline, S., & Somers, V. K. (2017). Sleep apnea: Types, mechanisms, and clinical cardiovascular consequences. *Journal of the American College of Cardiology*, *69*(7), 841–858. https://doi.org/10.1016/j.jacc.2016.11.069
13. Kingshott, R. (2017, noviembre 7). AASM clarifies hypopnea scoring criteria. *American Academy of Sleep Medicine – Association for Sleep Clinicians and Researchers*. https://aasm.org/aasm-clarifies-hypopnea-scoring-criteria/
14. Konopka, B. M., Lwow, F., Owczarz, M., & Łaczmański, Ł. (2018). Exploratory data analysis of a clinical study group: Development of a procedure for exploring multidimensional data. *PLOS ONE*, *13*(8), e0201950. https://doi.org/10.1371/journal.pone.0201950
15. Mayo Clinic. (s. f.). Sleep apnea - Symptoms and causes. https://www.mayoclinic.org/diseases-conditions/sleep-apnea/symptoms-causes/syc-20377631
16. Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). *Foundations of machine learning* (2.ª ed.). The MIT Press.
17. Nagarajan, G., & Dhinesh Babu, L. D. (2022). Missing data imputation on biomedical data using deeply learned clustering and L2 regularized regression based on symmetric uncertainty. *Artificial Intelligence in Medicine*, *123*, 102214. https://doi.org/10.1016/j.artmed.2021.102214
18. Organización Mundial de la Salud. (2024, marzo 1). Obesidad y sobrepeso. https://www.who.int/es/news-room/fact-sheets/detail/obesity-and-overweight

19. Park, C., Byun, J. I., Choi, S. H., Kim, K. T., Lee, J. S., Kim, D. K., & Seo, W. K. (2025). Machine learning classifier solving the problem of sleep stage imbalance between overnight sleep. *Biomedical Engineering Letters*, *15*, 513–523. https://doi.org/10.1007/s13534-025-00466-8
20. Patel, S. R., Larkin, E. K., & Redline, S. (2008). Shared genetic basis for obstructive sleep apnea and adiposity measures. *American Journal of Epidemiology*, *167*(9), 1006-1014. https://doi.org/10.1093/aje/kws342
21. Patiño-Pérez, D., Iñiguez-Muñoz, F., Ochoa-Flores, Á., Córdova-Aragundi, J., Castro-Carrasco, J., Luque-Letechi, A., & Munive-Mora, C. (2023). Estratificación para mejorar el rendimiento de una ANN en la detección de diabetes.
22. Rajkomar, A., Dean, J., & Kohane, I. S. (2019). Machine learning in medicine. *The New England Journal of Medicine*, *380*(14), 1347–1358. https://doi.org/10.1056/NEJMra1814259
23. Redline, S., Sotres-Alvarez, D., Loredo, J., Hall, M., Patel, S. R., Ramos, A., Shah, N., Ries, A., Arens, R., Barnhart, J., Youngblood, M., Zee, P., & Daviglus, M. L. (2014). Sleep-disordered breathing in Hispanic/Latino individuals of diverse backgrounds: The Hispanic Community Health Study/Study of Latinos. *American Journal of Respiratory and Critical Care Medicine*, *189*(3), 335-344. https://doi.org/10.1164/rccm.201309-1735oc
24. Scikit-learn developers. (s. f.-a). *GradientBoostingClassifier*. Scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html
25. Scikit-learn developers. (s. f.-b). *RandomForestClassifier*. Scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
26. Scikit-learn developers. (s. f.-c). *sklearn.model_selection.StratifiedKFold*. Scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html
27. Scikit-learn developers. (s. f.-d). *SVC*. Scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html
28. Shamout, F., Zhu, T., & Clifton, D. A. (2021). Machine learning for clinical outcome prediction. *IEEE Reviews in Biomedical Engineering*, *14*, 116–126. https://doi.org/10.1109/RBME.2020.3007816
29. Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2018). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE Journal of Biomedical and Health Informatics*, *22*(5), 1589–1604. https://doi.org/10.1109/JBHI.2017.2767063
30. Sidey-Gibbons, J. A. M., & Sidey-Gibbons, C. J. (2019). Machine learning in medicine: A practical introduction. *BMC Medical Research Methodology*, *19*(1), 64. https://doi.org/10.1186/s12874-019-0681-4
31. Sleep Heart Health Study Research Group. (1997). The Sleep Heart Health Study: Design, rationale, and methods. *Sleep*, *20*(12), 1077-1085. https://pubmed.ncbi.nlm.nih.gov/9493915/
32. Steyerberg, E. W., & Harrell, F. E., Jr. (2016). Prediction models need appropriate internal, internal-external, and external validation. *Journal of Clinical Epidemiology*, *69*, 245–247. https://doi.org/10.1016/j.jclinepi.2015.04.005

33. Vensel-Rundo, J. (2019). Obstructive sleep apnea basics. *Cleveland Clinic Journal of Medicine*, *86*(Suppl. 1), 2–9. https://doi.org/10.3949/CCJM.86.S1.02
34. Yathish, T., & Manjula, C. (2024). Risk assessment of obstructive sleep apnoea symptoms and its correlation with oral manifestations: A cross-sectional study. *Journal of Clinical and Diagnostic Research*, *18*(4), 9–12.
35. Zhang, G., Cui, L., Mueller, R., Tao, S., Kim, M., Rueschman, M., Mariani, S., Mobley, D., & Redline, S. (2018). The National Sleep Research Resource: Towards a sleep data commons. *Journal of the American Medical Informatics Association*, *25*(10), 1351-1358. https://doi.org/10.1093/jamia/ocy064
36. Zheng, A., & Casari, A. (2018). *Feature engineering for machine learning: Principles and techniques for data scientists*. O'Reilly Media.