PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

FACULTAD DE CIENCIAS E INGENIERÍA

PROPUESTA DE MEDIDAS PALIATIVAS Y DETECCIÓN DE PERSONAS FUMADORAS CON EL USO DE BIOSEÑALES Y MODELOS PREDICTIVOS

Tesis para obtener el título profesional de Ingeniero Industrial

AUTOR:

Jorge Alberto Uriol Lescano

ASESOR:

Ing. Walter Alejandro Silva Sotillo

Lima, Marzo, 2025

Informe de Similitud

Yo, WALTER ALEJANDRO SILVA SOTILLO

docente de la Facultad de CIENCIAS E INGENIERIA de la Pontificia Universidad Católica del Perú, asesor(a) de la tesis/el trabajo de investigación titulado

Propuesta de medidas paliativas y detección de personas fumadoras con el uso de bioseñales y modelos predictivos

del/de la autor(a)/ de los(as) autores(as)

Jorge Alberto Uriol Lescano,

dejo constancia de lo siguiente:

- El mencionado documento tiene un índice de puntuación de similitud de 20%. Así lo consigna el reporte de similitud emitido por el software *Turnitin* el 01/12/2024.
- He revisado con detalle dicho reporte y la Tesis o Trabajo de Suficiencia Profesional, y no se advierte indicios de plagio.
- Las citas a otros autores y sus respectivas referencias cumplen con las pautas académicas.

Lugar y fecha: Tampa, 1ro de Diciembre de 2024

Apellidos y nombres del asesor / de la asesora: SILVA SOTILLO, WALTER ALEJANDRO Paterno Materno, Nombre1 Nombre 2	
DNI: 09880013	Firma
ORCID: 0000-0003-3162-6340	

Resumen

La investigación analiza el impacto del tabaquismo en el Perú y propone un enfoque basado en el uso de bioseñales y modelos predictivos para la detección temprana de fumadores. Este trabajo se justifica por el alto costo económico y social del tabaquismo, que representa el 4% del PBI nacional y es responsable de 16,719 muertes anuales. La investigación busca identificar a fumadores y prevenir los efectos nocivos en su salud y en la de las personas expuestas al humo del tabaco.

Los objetivos incluyen diseñar un modelo predictivo eficaz y evaluar el impacto de las medidas propuestas en la calidad de vida y los costos en salud. Teóricamente, se sustenta en la minería de datos y el aprendizaje automático, empleando algoritmos como árboles de decisión, regresión logística y redes neuronales. Metodológicamente, se analizaron datos demográficos y biométricos utilizando técnicas de correlación y validación cruzada para garantizar la precisión y robustez del modelo.

Los resultados muestran que las técnicas predictivas permiten detectar fumadores con alto grado de precisión, posibilitando acciones tempranas para reducir el tabaquismo y sus consecuencias. Como conclusión principal, se proyecta que la implementación de estas herramientas podría disminuir la mortalidad por tabaquismo en un 30% y reducir los costos del sector salud generando un ahorro aproximado del 29%. Este enfoque ofrece una solución innovadora y efectiva al problema del tabaquismo en el contexto peruano.

Agradecimiento

Quiero expresar mi más profundo agradecimiento a todas las personas que, de manera directa o indirecta, hicieron posible que hoy culmine esta etapa tan importante de mi vida.
A mi familia, por estar a mi lado en todo momento, brindándome su apoyo incondicional y siendo mi principal fuente de fortaleza durante este camino. Su paciencia, amor y aliento me han permitido superar cada desafío que se presentó.

A mis profesores, quienes no solo me guiaron en el ámbito académico, sino que también me inculcaron la chispa de la curiosidad y el deseo de aprender continuamente. Sus enseñanzas trascienden las aulas y quedarán conmigo para siempre.

Y, por supuesto, a mis amigos, quienes han estado ahí en todo momento, alentándome a seguir adelante. Su apoyo, compañía y palabras de ánimo fueron fundamentales para mantenerme motivado y enfocado en mis metas.

A todos ustedes, gracias. Este logro no sería posible sin su presencia y contribución.

ÍNDICE

Índice de tabla

Índice de Gráficos

Introducción

En el período de la información, la minería de datos se ha convertido en una tecnologíavital que permite extraer ideas útiles de conjuntos de datos masivos. En este proceso se utilizan varias herramientas analíticas para encontrar patrones ocultos, tendencias y relaciones que pueden ser importantes para tomar decisiones. Sin embargo, este campo no está exento de desafíos y riesgos, lo que subraya la importancia de comprender a fondo sus fundamentos teóricos y las técnicas asociadas.

En el primer capítulo, se examina la definición de minería de datos, destacando los riesgos y desafíos inherentes a este proceso. Se examinan también las técnicas fundamentales empleadas en la minería de datos, desde las redes neuronales hasta algoritmos como las regresiones logísticas, árboles de decisión, máquinas de vectores de soporte,Extreme Gradient Boosting y Random Forest. Estos análisis detallados sientan las bases para entender cómo estas técnicas se aplican en diversos contextos, como se detalla en la sección sobre el empleo de la minería de datos.

En el segundo capítulo, se realiza una inmersión profunda en la situación actual relacionada con los fumadores en el contexto peruano. Se abordan aspectos como el fumador pasivo, el impacto del cigarrillo electrónico y el aumento de la mortalidad vinculado al tabaquismo. Además, se explora el impacto económico que los fumadores generan en el Perú, analizando enfoques de costos asociados al consumo de tabaco. El diagnóstico detallado de la problemática, incluyendo un análisis de causas utilizando herramientas como el diagrama causa-efecto y el diagrama de Pareto, proporciona una visión integral de la situación.

El tercer capítulo se centra en el procesamiento y análisis de datos, presentando una tabla detallada de datos, características relevantes y la construcción de estos para su posterior análisis. Se examina la correlación entre las variables, utilizando herramientas como la matriz de correlación de Pearson y la matriz de correlación de punto biserial.

En el cuarto capítulo, se propone un modelo predictivo para abordar la problemática identificada. Se seleccionan diversas técnicas de modelado, desde árboles de

decisión hasta redes neuronales, evaluando cuantitativamente sus resultados mediante pruebas de precisión.

El quinto capítulo evalúa los resultados obtenidos, estimando las implicancias económicas de la aplicación del modelo propuesto. Se analiza el mejor plan de acción, se proyecta el impacto esperado de la mejora en términos económicos y de calidad de vida. Por último, se muestran las conclusiones y recomendaciones centradas en los hallazgos de la investigación.

Este trabajo se posiciona como un aporte significativo para comprender, analizar y proponer soluciones innovadoras en la intersección entre la minería de datos y la problemática del tabaquismo en el contexto peruano.

Capítulo 1. Marco teórico

En el actual capítulo se elaborará descripciones acerca de la minería de datos, los modelos existentes para finalmente concluir con ejemplos de la aplicación de dichas técnicas.

1.1 ¿Qué es la minería de datos?

Actualmente podemos encontrar datos sobre una amplia gama de temas en diversos lugares. Sin embargo, el uso de métodos tradicionales hace imposible observar gran parte de estos datos, por lo que recurrimos a la minería de datos, que emplea una serie de técnicas, incluida la inteligencia artificial, para hallar patrones y vinculaciones dentro con el fin de crear modelos, o representaciones abstractas de la realidad.

La aplicación de las diversas técnicas tiene como objetivo el incorporar conocimientos obtenidos a través del análisis de estos datos para lograr tomar decisiones adecuadas basadas en las representaciones abstractas creadas.

La minería de datos es una fase crucial en el proceso integral de descubrimiento del conocimiento. Su objetivo principal es identificar patrones o modelos en los datos recopilados, evaluando si alguno de estos modelos puede generar resultados óptimos. Los algoritmos de minería de datos generalmente incluyen tres componentes clave: (Ballesteros, Iñiguez, & Velasco, 2018)

i) El modelo, que incluye parámetros determinados a raíz de la data de entrada.
ii) El criterio de preferencia, utilizado para evaluar y contrastar diferentes modelos alternativos.
iii) Los algoritmos de búsquedas.

1.1.1 Los riesgos y desafíos de la minería de datos

Uno de los riesgos más grandes que implica la minería de datos es el manejo de información potencialmente sensible o de identificación personal, en este

punto se encuentran las mayores preocupaciones pues la pérdida de los mismos o la posibilidad de pérdida puede afectar a miles de personas.

Por otra parte, también es importante mantener la confiablidad de los datos, la cual permite lograr un análisis correcto, completo, preciso y confiable. Debemos recordar que estos datos serán utilizados para tomar decisiones importantes y que se requerirán nuevas tecnologías para reunir datos del entorno de una las nuevas formas de modernas de datos. (Asencios, 2004)

1.2 Técnicas usadas por la minería de datos

Existen diversas técnicas utilizadas por la minería de datos, las cuales serán descritas, así como, las ventajas y desventajas de estas.

1.2.1 Redes neuronales

El funcionamiento del cerebro humano es muy complejo, este sistema está compuesto por redes neuronales interconectadas permiten realizar actividades superiores del ser humano como aprender, tomar decisiones, entre otras. Como cualquier sistema, estas redes neuronales biológicas "aprenden" de las informaciones que reciben y, bajo ciertas condiciones, muestran las capacidades de generalizar informaciones más allá de su entrenamiento. Esta característica especial es lo que intentan lograr las redes neuronales artificiales (RNA), es decir, las redes neuronales artificiales intentan imitar las funciones del cerebro humano.

Las redes neuronales artificiales (RNA) han mostrado su eficacia para abordar problemas complejos en los que los sistemas de computación tradicionales han enfrentado desafíos significativos durante mucho tiempo. Entre las aplicaciones exitosas de las RNA se encuentran los procesamientos de imágenes, el procesamiento de voz, la identificación de patrones, las planificaciones, las predicciones, los controles y la optimización.

Los tradicionales computadores funcionan basándose en la arquitectura de Von Neumann, procesando las informaciones de manera secuencial. En este tipo de computadoras, un único procesador maneja tanto las instrucciones

como los datos almacenados en la memoria, lo que se conoce como computación serial. El procesador accede y ejecuta las instrucciones de la memoria una por una.

Por otro lado, una red neuronal no sigue un proceso secuencial. En su lugar, responde en paralelo a varias entradas que se le dan, y cuando el resultado llega a la red se equilibra. Los conocimientos en una red neuronal no se almacenan en un único aprendizaje. Su poder reside en la topología y el número de conexiones entre neuronas.

Cuando se trata de aprender redes neuronales, esto se refiere a ajustar sus pesos de manera que su comportamiento global sea el más adecuado. Para lograrlo, se emplean diversos algoritmos, cada uno con variantes que se adaptan mejor según las condiciones y el contexto del problema.

Después del entrenamiento, la validación se realiza utilizando el conjunto de prueba, es decir, los datos que no se utilizaron en la fase de entrenamiento.

En la red neuronal artificial (RNA), el equivalente de las neuronas biológicas es un elemento de procesamiento llamado PE (process element). Este elemento de procesamiento recibe múltiples entradas, las combina y luego utiliza una función de transferencia para modificar el resultado. El valor de salida que se obtiene es producto de esta función de transferencia.

Las salidas de un PE se pueden conectar a las entradas de otros PE mediante conexiones ponderadas que reflejan la eficiencia de las sinapsis en esas conexiones. En términos gráficos, la estructura de la red neuronal se puede visualizar en el gráfico 1, donde se observan dos capas conectadas al entorno externo. La capa de entrada, que actúa como un buffer, representa los datos que ingresan a la red, mientras que la capa de salida, también un buffer, representa los resultados o salidas de la red. Las capas intermedias se denominan capas ocultas. (OCAMPO, GIRALDO, & ISAZA, 2006)

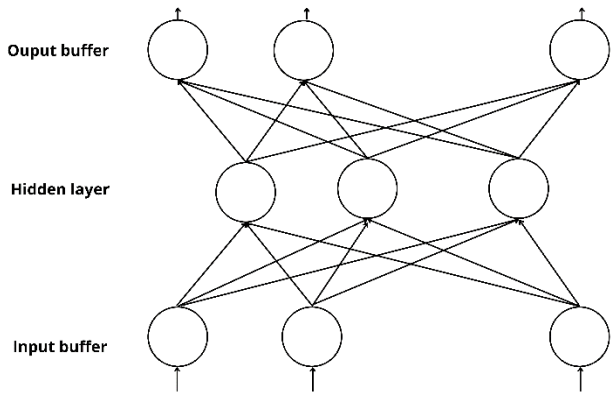

Gráfico 1 Arquitectura de las redes neuronales

Fuente: Escuela Superior de Ingeniería de Bilbao

1.2.1.1 Operación de una red neuronal artificial

Las estructuras topológicas de redes neuronales se describen como n-h-s, donde **n** representa los nodos de entrada, **h** las neuronas ocultas y **s** las neuronas de salida. Las neuronas ocultas reciben los valores de entrada (como se observa en la ecuación 01) X_j, con j=1…n, Estas neuronas ocultas procesan las entradas xj ponderadas por los pesos de W_{ij} y bias b_i después producen las salidas Y_i (salidas de cada neurona de la segunda capa), calculada como:

$$Y_{i=}\emptyset_i\left(\sum_{j=1}^{n} W_{ij}X_j + b_i\right) \quad \text{Con i =1,….,h}$$

Donde $\emptyset_i$ es la función de transferencia de la neurona i. Por último, la salida de las neuronas ocultas se pondera y suma, y se aplica una función de activación, lo que da como resultado las salidas finales de la red neuronal:

$$Y_{s=}\emptyset_s\left(\sum_{j=1}^{n} W_{si}Y_i + b_s\right) \text{ con s=1,…,g}$$

Los valores de salida dependerán de lo que se desea estimar.

1.2.1.2 Ventajas y desventajas redes neuronales

Ventajas de la red neuronal

- Capacidad de aprender: Las redes neuronales tienen las capacidades de aprender de los datos sin ser programadas. Lo que las convierte en herramientas poderosas para los reconocimientos de imágenes y los procesamientos del lenguaje natural.
- Capacidad de generalizar: Las redes neuronales pueden generalizar nuevos datos que son similares a los que fueron entrenados, esto resulta útil para procesos de previsión y la clasificación
- Robustez al ruido: Las redes neuronales son bastante resistentes al ruido, esto les permite ser ideales para el reconocimiento de imágenes, donde los datos pueden estar incompletos.

Desventajas de las redes neuronales

- Requiere una gran cantidad de datos: Esto puede resultar costoso y en ocasiones no llega a tener los resultados esperados por la falta de estos.
- Propenso al sobreajuste: Las redes neuronales pueden sobrepasarse fácilmente, lo que significa que pueden aprender material de aprendizaje demasiado bien y generalizar mal el material nuevo.
- Difícil de interpretar: Los resultados obtenidos de las redes neuronales llegan a ser difíciles de comprender, esto conlleva a no lograr entender la toma de decisiones de la red neuronal.

1.2.2 Regresión logística

La regresión logística fue desarrollada en la década de 1960 como una alternativa a la estimación por mínimos cuadrados ordinarios, comúnmente utilizada en modelos de regresión lineal. Su finalidad es valorar las probabilidades de un evento con base en un conjunto de variables independientes. En distinción de la regresión lineal, la regresión lineal se utiliza para predecir valores continuos, las regresiones logísticas se utilizan

cuando el resultado que deseamos predecir es una variable categórica o discreta. Su objetivo principal es establecer la relación entre un conjunto de variables independientes y una variable dependiente binaria.

Para ello, la regresión logística emplea una función logística que transforma un conjunto lineal de variables independientes en probabilidades. Este modelo describe el logaritmo del odds ratio como una función lineal de las variables independientes. La ecuación general de la regresión logística se expresa de la siguiente manera:

$$log(p\ /\ (1-p))\ =\ \beta 0\ +\ \beta 1x1\ +\ \beta 2x2 + \ldots + \beta nxn$$

Donde:

- P representa la probabilidad de ocurrencia del evento binario
- $\beta 0, \beta 1, \beta 2 \ldots \beta n$ son los coeficientes de regresión que se estiman a partir de los datos
- $x1, x2, \ldots xn$ son los valores de las variables independientes.

Una vez que se han calculado los coeficientes estimados se pueden predecir probabilidades de eventos para conjuntos de valores nuevos de las variables independientes. Esto permite evaluar la probabilidad de que el evento ocurra, dado un nuevo conjunto de datos, utilizando la ecuación de la regresión logística ajustada.

Este método es muy utilizado en medicina para predecir enfermedades, analizar factores de riesgo, diagnosticar y evasiva a tratamientos. (Puga, 2009)

1.2.2.1 Ventajas y desventajas regresión logística

Ventajas

- Fácil de entender e interpretar: La regresión logística es un modelo relativamente simple que es fácil de entender e

interpretar. Los resultados se pueden llegar a comunicar fácilmente a las partes interesadas.

- Eficiente para entrenar: Los modelos de regresión logística se pueden entrenar de manera rápida y eficiente, aunque el conjunto de datos sea grande.
- Versátil: La regresión logística tiene la capacidad de modelar una variedad de resultados binarios incluida la rotación de clientes, el diagnóstico médico, entre otros.
- Robustos frente valores atípicos: Los modelos de regresión logística suelen ser robustos frente a los valores atípicos.

Desventajas

- Linealidad: La regresión logística toma una vinculación lineal entre las variables independientes y el algoritmo de la razón de probabilidades. Si esta relación es no lineal las predicciones pueden ser imprecisas.
- Independencia de observaciones: La regresión logística asume que las observaciones son independientes entre sí. Si existe correlación entre las observaciones, es necesario utilizar otras técnicas de análisis.
- Ausencia de multicolinealidad: Esto sucede cuando existe una fuerte correlación entre las variables independientes. Esto dificulta las interpretaciones de la regresión y dar lugar a resultados sesgados.
- Tamaño de la muestra: Debe ser suficientemente grande el tamaño de la muestra para lograr confiables resultados.
- Ausencia de causalidad: Puede identificar asociaciones entre las variables independientes y la variable dependiente, pero no puede establecer causalidad.

1.2.3 Máquinas de Vectores de soporte

Las máquinas de vectores de soporte (SVM) son algoritmos de aprendizajes supervisados extensamente empleado en problemáticas de clasificación y regresión. Este enfoque fue propuesto por Vladimir Vapnik y su equipo en

1990, y ha encontrado aplicaciones significativas en campos como la medicina, la visión por computadora y la minería de datos.

El concepto central de las SVM es la identificación de un hiperplano en un espacio de alta dimensión que separa de manera óptima muestras de diferentes categorías. En otras palabras, el propósito es hallar un límite de decisión que maximice la distancia entre las clases. El hiperplano se define empleando vectores de soporte, que son los puntos más próximos a este límite de decisión.

La SVM se basa en maximizar el margen de la distancia entre el hiperplano y los puntos de datos más cercano, esto ayuda a obtener un límite de decisión más generalizado y menos propenso al sobreajuste. (Velásquez, Olaya1, & Franco, 2010)

1.2.3.1 Ventajas y desventajas SVM

Ventajas del SVM:

- Eficiencia en espacios de alta dimensión: Las SVM son efectivas en espacios de alta dimensión, es adecuado cuando existen muchas variables.
- Manejo de datos no lineales: Las SVM puede manejar eficientemente datos que no son linealmente separables mediante el uso de funciones de kernel.
- Regularización incorporada: Las SVM incorporan un término de regularización para controlar el sobreajuste.
- Manejo de datos desbalanceados: SVM puede tratar eficazmente conjuntos de datos desbalanceados, donde una clase tiene muchos más ejemplos que la otra, mediante el ajuste de pesos en función de su importancia relativa.

Desventaja del SVM:

- Sensibilizada a la escala de los datos: El modelo puede ser sensible a la escala de las variables, por ello se recomiendo escalar los datos antes.
- Selección de parámetros: Las SVM tiene parámetros que deben ser ajustados, como el tipo de kernel y los parámetros asociados. La selección adecuada requiere validación cruzada u optimización.

1.2.3.2 Formulación Matemática de las SVM

El modelo:

Dada una serie de tiempo y_t con regresores x_t, y contando con un conjunto de D ejemplos representativos, una SVM permite aproximar esta relación mediante la siguiente función:

$$y_t = b + \sum_{d=1}^{D} w_d Xk(x_t, x_d) \ldots\ldots (1)$$

Donde b es una constante y w_d son los coeficientes de ponderación de la función de núcleo k. De este modo, una SVM representa combinaciones lineales del mapeo de x_t en un espacio de particularidades altamente no lineal, definido por los puntosx_d y la función de transformación no lineal.

Estimación:

La estimación se realiza a través de la minimización de la función de riesgo regularizado $R(C, \epsilon)$, que se define como:

$$R(C, \epsilon) = C\frac{1}{D}\sum_{d=1}^{D} L_\epsilon(y_d, \acute{y_d}) + \frac{1}{2}\sum_{d=1}^{D} {w_d}^2 \ldots\ldots (2)$$

Donde el primer término evalúa el error empírico entre el modelo y los datos, mientras que el segundo término corresponde a la parte de regularización, la cual depende exclusivamente de los pesos w_d. La constante de regulación C permite ajustar la importancia relativa de estos dos componentes. A medida que C se aproxima a cero, la magnitud de la función $R(C, \epsilon)$ depende únicamente de w_d, independientemente de cómo se ajusten los datos, lo que conduce a una disminución de los pesos w_d tanto como sea posible. L_ϵ es la función definida como:

$$L_\epsilon(y_d, \acute{y_d}) = \begin{cases} \left|(y_d - \acute{y}_d)\right| & para\ \left|(y_d - \acute{y}_d)\right| > \epsilon \\ 0 & \end{cases} \ldots\ldots (3)$$

Donde la constante ϵ representa la precisión deseada y define el radio de un margen (o tubo) dentro del cual el error es considerado como cero.

Funciones de núcleo (kernels)

En la ecuación (1), $k(x_t, x_d)$ es una función de núcleo que permite mapear el punto x a un espacio de alta dimensionalidad parametrizado por los puntos x_d . existen varias funciones de núcleo que se utilizan comúnmente.

- Lineal: $x\ x_d$
- Polinomial: $(a_1 x x_d + a_2)^{\,d}$
- Gaussiana o RBF: $exp\left(\left[-\frac{1}{{a_1}^2}\right] x(x - x_d)^{\,2}\right)$
- Exponencial: $exp\left(\left[-\frac{1}{{a_1}^2}\right] x(x - x_d)\right)$
- Perceptron multicapa: $tanh\ (a_1 x x_d + a_2)$

Las constantes a_1 y a_2 dependen del problema específico y deben ser ajustadas por el modelador.

La metodología para seleccionar estas constantes varía según el problema que se desea analizar, y no existen métodos heurísticos establecidos para determinar sus valores de manera general.

1.2.4 Árboles de decisión

El modelo de árbol de decisión son algoritmos de aprendizajes supervisados que se emplea tanto en problemáticas de clasificación como de regresión. Este modelo está estructurado en forma de un árbol de decisiones, donde cada nodo interno corresponde a una prueba o decisión centrada en una específica característica de las informaciones, cada rama expresa el resultado de dicha prueba, y cada hoja termina con predicciones, tal como se evidencia en el gráfico 2.

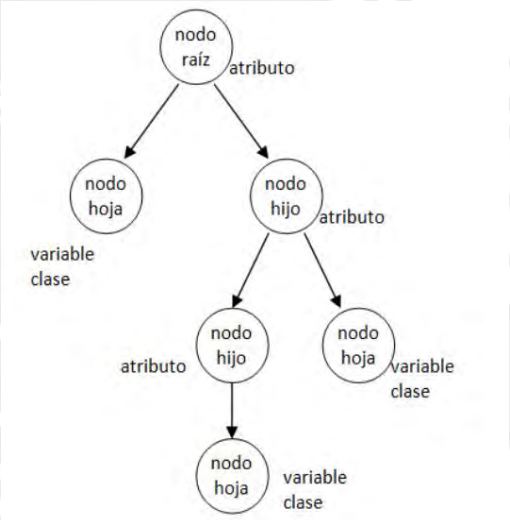

Gráfico 2 Estructura de un árbol de decisión

Fuente: Instituto de Ciencias de la Salud, Universidad Veracruzana

Los algoritmos para generar árboles de decisión se desarrollan en dos fases: inducción de árboles y clasificación. En el primer paso, se construye un árbol basado en los datos de entrenamiento proporcionados. En este paso, cada nodo interno del árbol se vincula con un atributo de prueba y el conjunto de entrenamiento para ese nodo se divide en función de los posibles valores de ese atributo. Un árbol se construye creando primero un nodo raíz,

escogiendo atributos de prueba y dividiendo el conjunto de entrenamiento en dos o más subconjuntos, creando cada partición un nuevo nodo.

En el segundo paso, el algoritmo utiliza el árbol creado previamente para clasificar cada nuevo objeto. Este procedimiento involucra recorrer el árbol desde el nodo raíz hasta llegar a la hoja que brindan las predicciones finales. (Martínez, y otros, 2009)

1.2.4.1 Ventajas y desventajas de los árboles de decisiones

Ventajas de árboles de decisión

- Interpretabilidad**:** Los árboles de decisión son sencillos de comprender e interpretar, ya que pueden visualizarse y seguirse de manera intuitiva.
- Manejo de datos mixtos: Los árboles de decisión pueden trabajar con datos tanto numéricos como categóricos sin requerir un preprocesamiento previo.
- Manejo de relaciones no lineales: Estos modelos captan las vinculaciones no lineales entre las variables.

Desventaja de árboles de decisión

- Tendencia al sobreajuste: Al crear particiones muy específicas, pueden llevar al sobreajuste del modelo.
- Sensibilidad a pequeñas variaciones en los datos: Alteraciones pequeñas en las informaciones pueden generar resultados significativamente diferentes.
- Dificultad para capturar relaciones complejas: Pueden presentar dificultades para modelar complejas relaciones entre las variables

1.2.5 Random forest

Random Forest son algoritmos de aprendizajes automáticos que mezcla múltiples árboles de decisión independientes para realizar tareas de clasificaciones y regresiones.

Esquematizando, el proceso de funcionamiento del algoritmo en lo siguiente:

1. Se dividen los datos en dos conjuntos: uno para entrenamiento y otro para prueba.

2. Se construye un bosque aleatorio utilizando el conjunto de entrenamiento, siguiendo estos pasos para cada árbol:

- Se seleccionan aleatoriamente "n" datos con repetición del conjunto de entrenamiento.
- Se entrena un árbol de decisión utilizando esta muestra.
- Si existen M entradas (inputs), se selecciona aleatoriamente un número m de ellas para utilizar en la decisión en cada nodo del árbol.
- Se iteran todos los posibles valores de cada entrada seleccionada en m.
- Se continúa dividiendo los nodos en dos subnodos hasta alcanzar el tamaño deseado de nodos, obteniendo así el árbol i

3. Finalmente, se utiliza el grupo de prueba para realizar predicciones y calcular los rendimientos del modelo.

1.2.5.1 Ventajas y desventajas de Random Forest

Ventajas de Random Forest

- Precisión: Los modelos de Random Forest tienden a tener una alta precisión debido a que en estos se combinan múltiples árboles independientes.
- Robustez frente a datos faltantes: Random Forest puede manejar datos ruidosos y faltantes, esto se debe a la combinación de múltiples árboles lo que reduce el impacto de los valores atípicos

- Importancia de características: Random Forest brinda medidas de la importancia de cada particularidad en los procesos de clasificaciones o regresiones.
    - Eficiencia en grandes conjuntos de datos: Puede emplear conjuntos grandes de datos con varias particularidades.

Desventajas de Random Forest

- Mayor complejidad: Random Forest puede ser más complejo que solo utilizar árbol de decisión, ya que este implica la combinación de varios.
- Ajuste de hiperparámetros. Random Forest presenta muchos hiperparámetros que se deben ajustar los números de árboles y profundidades máximas del árbol. La selección alterará el rendimiento del modelo.

1.2.5.2 Algoritmo Random forest

Cada árbol generado depende de un conjunto de variables aleatorias. Más formalmente, dado un vector aleatorio $X = (X_1, \dots, X_P)^T$, que representa la respuesta con valores reales, se asume una distribución conjunta desconocida $(P_{XY}(X, Y))$ El propósito es hallar una función de predicción f(x) que prediga el valor de Y. La función de predicción se determina mediante la función de pérdida $L(Y, f(X))$ Esta función de predicción se define minimizando el valor esperado de la pérdida :

$$E_{XY}(L(Y, f(X)))$$

Donde los subíndices indican la expectativa con respecto a la distribución conjunta de X e Y. De manera intuitiva, $L(Y, f(X))$ mide qué tan cerca está F(x) de Y; penaliza aquellos valores de $f(X)$ que se desvían de Y. Las opciones comunes para L incluyen la pérdida de error cuadrático, $(L(Y, f(X)) = (Y - f(X))^2)$, que se utiliza en problemas de regresión, y la pérdida de cero-uno, que se emplea en problemas de clasificación.

$$L(Y, f(X)) = I(Y\ 6 =\ f\ (X)) = \begin{cases} 0\ si\ Y = f(x) \\ 1\ en\ caso\ contrario \end{cases}$$

Resulta que minimizar *EXY* (*L*(*Y*, *f* (*X*)))para la pérdida de error cuadrático proporciona la esperanza condicional.

$$f\,(x) = E(Y|X = x)$$

De lo contrario, también conocida como la función de regresión. En el caso de la clasificación, si el conjunto de posibles valores de Y se denota como y, minimizar EXY (L(Y, f(X))) para la pérdida de cero-uno produce la función f que maximiza la probabilidad condicional, es decir:

$$f\,(x) = \textit{arg}\max\ p(Y=y|\ X=x)$$

Los ensambles construyen f en términos de una colección llamada "base learners" y esta base learners se combinan para dar el ensamble predictor f(x). En regresión, los base learners se promedian. (Cutler, Cutler, & Stevens, 2014)

$$(f(x) = (\tfrac{1}{J}) \sum_{j=1}^{J} h_j(x),$$

1.2.6 Extreme Gradient Boosting

El algoritmo Extreme Gradient Boosting (XGBoost) es una técnica de aprendizaje supervisado basada en árboles de decisión y se considera una evolución de los algoritmos tradicionales, como se observa en el gráfico 3.

Gráfico 3 Evolución de algoritmos basados en árboles de decisión

Fuente: Ingeniería Investigación y Tecnología volumen XXI (número 3)

1.2.6.1 Características de XG boost

Se puede definir como un ensamblaje secuencial de árboles de decisión, conocido como "Classification and Regression Trees" (CART). En este enfoque, los árboles de decisión se agregan de manera secuencial, y el modelo se encarga de aprender de los errores cometidos en las predicciones anteriores con cada árbol agregado, hasta que ya no sea posible corregir más ese error. Este proceso se conoce como "gradiente descendente", como se muestra en el gráfico 4.

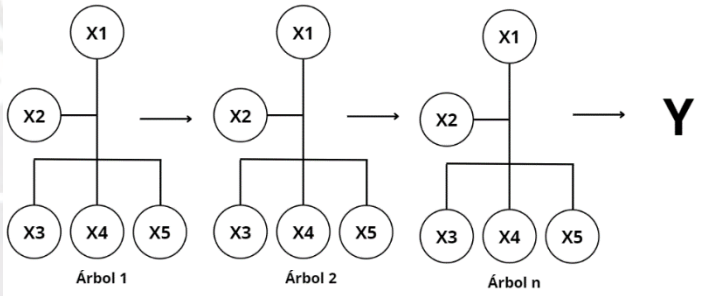

Gráfico 4 Algoritmo XG Boost

Fuente: Ingeniería Investigación y Tecnología volumen XXI (número 3)

La diferencia principal entre el algoritmo XGBoost y Random Forest es que en Random Forest se define la extensión de los árboles de antemano, mientras que XGBoost permite que los árboles se extiendan hasta su máxima profundidad. Además, XGBoost utiliza procesamiento en paralelo, incorpora poda de árboles, maneja valores perdidos de manera eficiente y emplea regularización para prevenir el sobreajuste o sesgo en el modelo.

1.2.6.2 Algoritmo XG boost

a) Primero, se construye un árbol inicial F_0 para predecir la variable objetivo “y”, donde el resultado se asocia con un residual $(y - F_0)$.

b) Luego, se genera un nuevo árbol H_1, el cual, como se explicó anteriormente, ajusta el error del paso previo.

c) Los resultados de F_0 y H_1 se combinan para obtener el árbol F_1, donde el error cuadrático medio es menor que el de F_0. Esto se expresa como:

$$F_1(x) < -F_0(x) + F_1(x)$$

d) Este proceso se repite iterativamente hasta que el error se minimiza lo máximo posible de la siguiente manera:(Chen & Guestrin, 2016)

$$F_m(x) < -F_{m-1}(x) + h_m(x)$$

1.2.6.3 Ventajas y desventajas del algoritmo XG Boost

Ventajas:

- Puede manejar grandes volúmenes de datos con múltiples variables.
- Puede gestionar valores perdidos de manera eficiente.
- Los resultados obtenidos son altamente precisos.
- Ofrece buena velocidad de ejecución.

Desventajas:

- Puede llegar a consumir muchos recursos computacionales en bases de datos grandes.
- Es necesario ajustar los parámetros del algoritmo para minimizar el error de precisión y evitar el sobreajuste del modelo.

- Solo funciona con vectores numéricos, por lo que se requieren transformaciones previas en caso de tener variables de otro tipo.

1.2.6 Cuadro resumen

En la tabla 01 se muestra un resumen con las diferentes ventajas y desventajas propias de cada modelo antes estudiado.

Tabla 1 Ventajas y desventajas modelos predictivos

Redes Neuronales	• Capacidad de aprender • Capacidad de generalizar Robustez al ruido	• Requiere una gran cantidad de datos • Propenso al sobreajuste
Regresión logística	• Fácil de entender • Eficiente para entrenar • Versátil con los datos • Robustos frente valores atípicos	• Linealidad • Independencia de observaciones • Ausencia de multicolinealidad • Tamaño de muestra grande • Ausencia de causalidad
Máquina de Vectores de soporte	• Eficiencia en espacios de alta dimensión • Manejo de datos no lineales Regularización incorporada • Manejo de datos desbalanceados	• Sensibilidad a la escala de los datos • Selección de parámetros dificultosa

Árboles de decisión	● Interpretabilidad sencilla ● Manejo de datos mixtos ● Manejo de relaciones no lineales	● Tendencia al sobreajuste Sensibilidad a pequeñas variaciones ● Dificultad para capturar relaciones complejas
Algoritmo random forest	● Precisión alta Robustez frente a datos faltantes ● Eficiencia frente a grandes grupos de datos	● Mayor complejidad ● Parámetros difíciles de configurar
Extreme Gradient Boosting	● Maneja grandes volúmenes de datos ● Manejo de datos perdidos ● Resultados precisos Buena velocidad de ejecución	● Puede llegar a consumir muchos recursos computacionales ● Parámetros dificultosos para configurar ● Solo trabaja con vector numéricos

Fuente: Elaboración propia

1.3 Aplicaciones de la minería de datos

En la siguiente parte se mostrarán algunos ejemplos de aplicaciones de estos modelos predictivos, en varios campos de estudio

1.3.1 Redes neuronales

Un ejemplo de aplicación de redes neuronales es el caso de las redes Backpropagation. En 1987, Sejnowski y Rosenberg lograron un gran avance

con su sistema llamado NetTalk, que tenía la capacidad de convertir texto en inglés en una voz altamente inteligible. Durante la fase de entrenamiento, la voz generada por el sistema recordaba los sonidos de un niño en diferentes etapas de aprendizaje. (Olabe, 2003)

1.3.2 Regresión logística

Un ejemplo de aplicación de la regresión logística se llevó a cabo en pacientes con diabetes mellitus, en un estudio realizado por el Instituto Nacional de Endocrinología. Esta aplicación consistió en una intervención educativa dirigida a los proveedores de salud en la comunidad. La población del estudio incluyó a 435 pacientes, y la intervención se desarrolló a lo largo de 5 años. Los pacientes fueron atendidos en dos policlínicos: 226 en el policlínico "Plaza de la Revolución", que funcionó como grupo experimental, y 209 en el policlínico "Héroes del Moncada", que actuó como grupo de control. Al finalizar los 5 años, se evaluó si el programa implementado generaba cambios en parámetros como el control metabólico y la aparición de complicaciones, entre otros. Uno de los resultados esperados fue relacionado en la disminución de complicaciones agudas, lo cual no tuvo el comportamiento esperado, ya que se observó hipoglicemia, por ello, se decidió analizar si esto obtuvo relación con otras variables mediante regresión logística, el resultado de la regresión reafirmó que la pertenencia a alguno de los policlínicos de salud no constituyó algún efecto para en la aparición de hipoglucemia. (Alonso & Padilla, 2001)

1.3.3 Máquinas de vectores de soporte

La revista Iberoamericana de Automática e Informática Industrial realizó una aplicación de las Máquinas de Vectores de Soporte (SVM) para el diagnóstico clínico de enfermedades como el Parkinson y el temblor esencial. Estas patologías, relacionadas con trastornos del movimiento, afectan significativamente la calidad de vida de quienes las padecen, causando discapacidad física, deterioro neurológico y, en muchos casos, exclusión social. Dado que es crucial identificar e iniciar tratamientos en las primeras etapas de los síntomas, se utilizó el enfoque de SVM para predecir

y diagnosticar estos casos de manera temprana. (González, Barrientos, Toapanta, & Cerro, 2017)

1.3.4 Árboles de decisión

Un ejemplo de aplicación de árboles de decisión se llevó a cabo en la Universidad Tecnológica de Pereira, en Colombia, por un grupo de ingenieros. Este proyecto buscaba aplicar el aprendizaje automático en el estudio de las variables del modelo de indicadores de gestión de las universidades públicas. El objetivo principal era mejorar la distribución de los recursos, ya que en Colombia no existe un modelo eficaz para su asignación. Actualmente, los recursos se distribuyen basándose en la medición de ciertos indicadores, que en algunos casos pueden proporcionar resultados erróneos, dificultando así el aumento de recursos para ciertas universidades. (Chávez, Mora, & Montoya, 2013

1.3.5 Random Forest

Los modelos de aprendizaje automático no solo tienen aplicaciones en el sector salud, sino también en el ámbito empresarial. Un ejemplo de ello se encuentra en el artículo publicado en Ingeniería, Inversión y Tecnología, volumen XXI. En este estudio, se utilizaron los algoritmos Random Forest y XGBoost en una base de datos de solicitudes de tarjetas de crédito. Un factor crítico para las instituciones financieras es identificar a los solicitantes a quienes otorgar una tarjeta de crédito, ya que reconocer a los usuarios con menores riesgos es fundamental para el negocio. En este contexto, se analizaron las variables más importantes, como si el usuario había solicitado previamente una tarjeta de crédito, para tomar decisiones más informadas. (Espinosa-Zúñiga, 2020)

1.3.6 Xgboost

En un paper publicado por IOP Conference Series: Materials Science and Engineering se utilizó el algoritmo XGboost para predecir la falla de rodamientos, el rodamiento es una de las partes que se dañan fácilmente en la rotación maquinaria, representando alrededor del 30% de las fallas. Este estudio recopiló datos de vibración del rodamiento a través de un sensor.

Con esto se logró un avance para identificar cuándo ocurrirán estas fallas, sin embargo, aún existen algunos impedimentos al momento de aplicarlo. (Zhang, Li, & Jiao, 2019)

1.3.7 Lasso

Un ejemplo de aplicación de Lasso se realizó en la Universidad de Concepción, utilizando un método de ajuste para seleccionar variables para predecir el riesgo de disfunción motora en adultos mayores de la ciudad de Valdivia. A través del uso de Lasso y otros métodos predictivos, se concluyó que solo es necesario identificar ciertas variables clave, lo que podría facilitar su implementación en consultas médicas convencionales. (FRITZ)

Capítulo 2. Descripción y análisis de la situación actual

Este capítulo abarcará el consumo de tabaco y su impacto en la realidad social, económica y salud del Perú, con un especial énfasis en ell impacto económico que conlleva.

2.1 Contexto de los fumadores en la actualidad peruana

El consumo de tabaco es un hábito perjudicial para la salud, y según la Organización Mundial de la Salud (OMS), provoca la muerte de aproximadamente 8 millones de personas al año. En Perú, el tabaquismo también es un problema significativo. De acuerdo con un estudio publicado por Andina en el año 2019, esta adicción representa una preocupación considerable en la realidad peruana. (Andina, 2019).

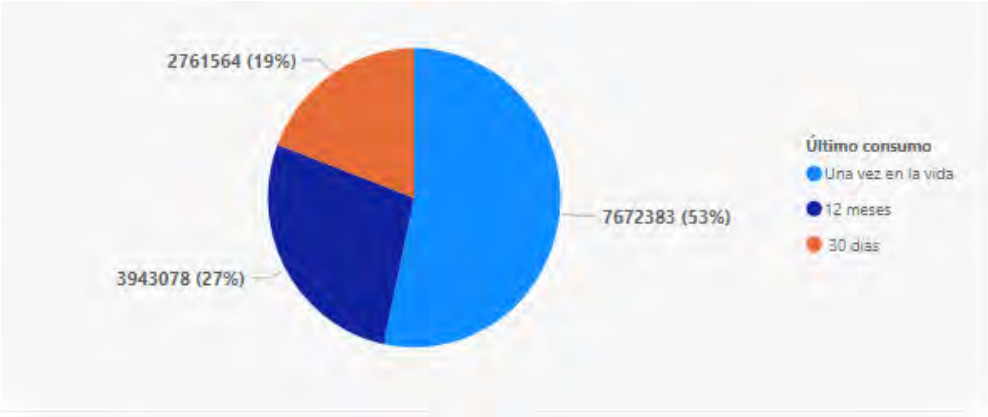

Gráfico 5 Consumo de tabaco

Fuente: Andina

En el gráfico 5 se muestra la cantidad y porcentaje de personas que consumen tabaco en Perú. En el mismo año, el Instituto Nacional de Enfermedades Neoplásicas (INEN) reportó que aproximadamente 167,000 peruanos fallecen anualmente debido a diversas enfermedades vinculadas con el consumo de tabaco, incluyendo el cáncer. Según un Boletín Epidemiológico publicado por el Ministerio de Salud, el cáncer es la segunda causa de muerte en el país.

El INEN también señala que la edad promedio del primer consumo de tabaco es de aproximadamente 18.4 años. Es en esta etapa donde muchas personas comienzan a fumar, y dependiendo de factores individuales, pueden desarrollar adicción al tabaco, especialmente a la nicotina. Esta sustancia es considerada tan adictiva como la heroína y la cocaína. Algunos de los efectos que la nicotina y otros químicos presentes en los cigarrillos tienen en el ser humano son los siguientes:

- La nicotina produce una sensación placentera y reduce las sensaciones desagradables, lo que genera la necesidad de continuar consumiéndola. Esta sustancia afecta las reacciones químicas en el cerebro y el SNC, afecta el estado de ánimo e inunda los circuitos cerebrales con dopamina, un neurotransmisor que está estrechamente vinculado con las sensaciones de placer.
- La nicotina llega al cerebro en cuestión de segundos dando las sensaciones placenteras antes mencionadas, luego de ello sus efectos desaparecen al cabo de un par de minutos, esto provoca que el sujeto comience a sentirse irritado y

tenso. Esta es la característica que ocasiona que la persona comience a sentir la necesidad de consumir tabaco nuevamente. Luego, la persona vuelve a consumir tabaco para no tener esas sensaciones desagradables y el ciclo se repite.

- También a medida que el tiempo pase la cantidad de tabaco que una persona consume comienza a aumentar pues las primeras dosis ya no tienen el mismo efecto placentero sobre él (Luis Pinillos A1, 2005).

Algunos síntomas que pueden experimentar después de dejar de consumir tabaco son los siguientes (médicos de la Sociedad Americana Contra El Cáncer, 2022):

- Mareos
- Depresión
- Sentimientos de frustración
- Ansiedad
- Intranquilidad
- Dolores de cabeza
- Cansancio
- Incremento de apetito
- Incremento de peso
- Ritmo cardíaco lento
- Estreñimientos y gases
- Opresión en el pecho
- Tos, boca seca dolor de garganta y goteo nasal

2.1.1 El fumador pasivo

Según la Organización Mundial de la Salud (OMS) y la Asociación Española Contra el Cáncer, un fumador activo inhala aproximadamente el 15% del humo del tabaco, mientras que el 85% restante es expulsado, lo que pone en riesgo a las personas que lo rodean. Las evidencias científicas muestran que el humo del tabaco causa varias enfermedades como cáncer, enfermedades respiratorias y diabetes. Asimismo, el humo del tabaco ha sido identificado como uno de los contaminantes principales en el hogar y el

entorno laboral. Los fumadores pasivos están expuestos a este humo equivalente a inhalar dos o tres cigarrillos.

Debido a la inhalación de numerosos componentes químicos nocivos presentes en el humo del cigarrillo, como alquitrán, monóxido de carbono, nicotina, arsénico, cadmio, cloruro de vinilo y cromo, las personas no fumadoras que están expuestas a estos compuestos tienen mayor probabilidad de desarrollar enfermedades respiratorias o cancerígenas en comparación con los fumadores activos.

Esta problemática, según el Ministerio de Salud, durante los últimos dos años, los especialistas de la Dirección de Control y Seguimiento de DIGESA han revisado el 20,6% de los locales donde se detectó tabaquismo. Para ello se toman medidas para disminuir y controlar el consumo de tabaco en lugares públicos. (MINSA, 2018)

Es importante mencionar que el desconocimiento sobre lo diferentes problemas que puede ocasionar el consumo de tabaco también es uno de los motivos primordiales por el cual el tabaquismo sigue aumentando, en la tabla 02 se puede observar el conocimiento de las enfermedades que el tabaquismo puede causar, en personas fumadoras, no fumadoras y los ex fumadores, de los cuales ctenían mayores conocimientos los fumadores y ex fumadores respectivamente.(Enrique Ruiz Mori, 2016)

Tabla 2 Conocimiento de enfermedades relacionadas al tabaquismo

	Fumadores			No fumadores			Ex Fumadores		
Conocimiento	Género		Total	Género		Total	Género		Total
	F	M		F	M		F	M	
Cáncer de Pulmón	173	187	360 (48,3 %)	116	170	286 (36,9%)	183	222	405 (53,8%)
Infarto al Miocardio	79	72	151 (20,3 %)	40	66	106 (13,6%)	54	83	137 (18,2%)
Enfisema Pulmonar	58	69	127 (17,0 %)	37	58	95 (12,2%)	46	62	108 (14,3%)
Impotencia Sexual	43	63	106 (14,2 %)	28	46	74 (9,5%)	30	57	87 (11,5%)

Fuente:9 Horiz Med 2016

2.1.2 El cigarrillo electrónico

El cigarrillo electrónico es otro de los factores que incita al usuario a caer en la adicción de consumo de la nicotina una de las principales razones por las cuales las personas comienzan a consumirlo según el gráfico 06 que resume las diferentes respuestas a esta pregunta es la de que el cigarrillo electrónico resulta ser menos dañino.

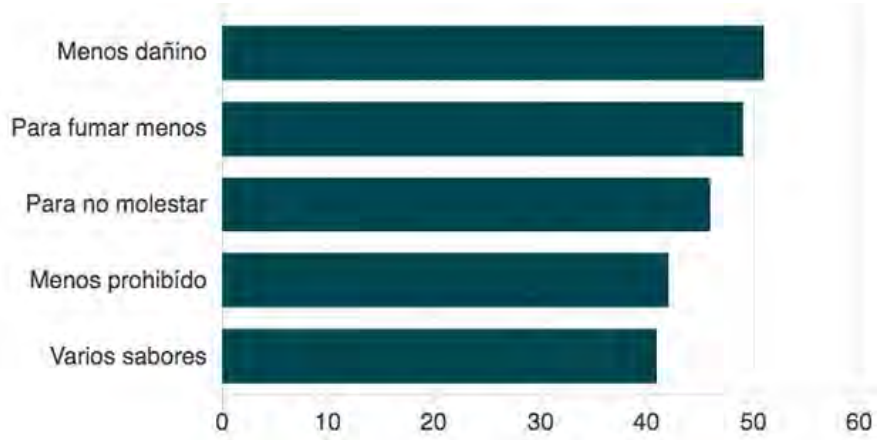

Gráfico 6 Respuesta razones para consumir cigarrillo electrónico

Fuente: Kantar Emst & Young analytics

A pesar de que el cigarrillo electrónico es menos dañino, esto puede incitar a las personas que lo consumen a caer en el consumo del cigarrillo convencional.

Según un estudio realizado por la Revista Médica Herediana (Rev Med Hered, 2020) los jóvenes empiezan a utilizar el cigarrillo electrónico desde los 15 años lo cual es preocupante y puede llevar a que en el futuro estas personas se conviertan en fumadores potenciales. (Núñez, 2019)

2.2 Aumento de la mortalidad debido a fumar

En primer lugar, es importante recalcar los problemas que surgen debido excesivo consumo de tabaco, los cuales pueden llevarnos a la muerte, algunos de estos son:

- Bronquitis crónica
- Enfisema pulmonar
- Cáncer de pulmón

- Hipertensión arterial
- Enfermedad coronaria
- Accidente cerebrovascular
- Úlcera gastrointestinal
- Gastritis crónica
- Cáncer de laringe, bucofaríngeo renal

En el Perú, la enfermedad más frecuente originada por el tabaquismo es la enfermedad pulmonar obstructiva crónica (EPOC). Esta es un padecimiento pulmonar que causa dificultad para respirar debido a la inflamación y daño a las vías respiratorias y los pulmones. Aunque el EPOC suele detectarse en etapas tempranas, no tiene cura, lo que obliga al paciente a asumir gastos adicionales y a replantear su estilo de vida actual.(Instituto Nacional de Estadística e Informática, 2023)

Estos son algunos de los problemas más graves que pueden provocar la muerte. Con respecto, la Organización Mundial de la Salud destacó en 2019 el enorme número de muertes por enfermedades pulmonares directamente relacionadas con el consumo de tabaco.

Más del 40% de los decesos vinculados con el tabaco están asociados con enfermedades pulmonares como el cáncer y diversas enfermedades respiratorias. Por este motivo, la OMS hizo un llamado a intensificar las medidas para proteger a las personas que están expuestas de manera continua al tabaco.

El tabaco es responsable de la muerte de más de ocho millones de personas cada año, de las cuales más de siete millones fallecen por el consumo directo, y aproximadamente 1.2 millones mueren debido a la exposición al humo del tabaco.

Se puede ver que tanto impacto tiene analizando el nivel de mortalidad, por ejemplo, en un estudio realizado en comunidades autónomas de España sobre la mortalidad podemos observar en el gráfico 07 la relación que existe entre la edad y posibles consecuencias que tiene. También el género influye pues en muchos lugares consumen más varones que mujeres, aunque también existen lugares donde esa predominancia es inversa (JuliaReya & LeonorVarela-Lemaa, 2017)

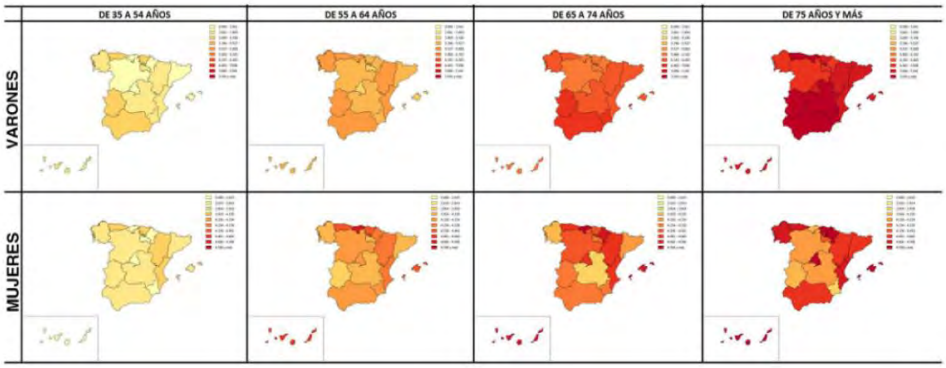

Gráfico 7 Tasa de mortalidad (Género/Edad)
Fuente: Universidad de Santiago de Compostela

En este caso los varones son los que más consumen tabaco y a mayor edad, mayor es el índice de mortalidad relacionado con el tabaquismo.

En este sentido, aún el tabaquismo sigue siendo uno de los factores más grandes que influyen en la mortalidad. La OMS también atribuye el tabaquismo como uno de los factores que podría influir en la mortalidad por COVID 19, también mencionó que es necesario evaluar los índices de mortalidad por tabaquismo en el mundo. Pues esto se relaciona directamente con el cáncer y esta es el principal motivo de muerte en el mundo.

El consumo crónico de tabaco, especialmente a través de la vía intrapulmonar (fumando), es altamente tóxico y provoca numerosas enfermedades, así como un elevado número de muertes prematuras. En los países desarrollados, es la principal causa de mortalidad prevenible.

La mortalidad entre los fumadores es mayor que la de los no fumadores siendo casi el doble. Se puede afirmar que, si no dejan de fumar, de cada cuatro fumadores, uno perderá aproximadamente entre 10 y 30 años de vida, otro perderá alrededor de 7 años, aunque su calidad de vida probablemente será muy deficiente. Los otros dos fumadores no verán una reducción significativa en su expectativa de vida, aunque la calidad de esta podría verse afectada. En promedio, los fumadores presentan una reducción de entre 5 y 8 años en su esperanza de vida. Además, se estima que un fumador resta entre 5 y 6 minutos de vida por cada cigarrillo fumado, lo que equivale, aproximadamente, al tiempo que se tarda en consumirlo.

En el gráfico 08 se muestra la relación entre el desarrollo de un país y la mortalidad asociada al tabaquismo, la cual, según el estudio, tiende a aumentar a medida que el país avanza en su desarrollo. Un patrón similar se observa en países que se encuentran en proceso de transición.

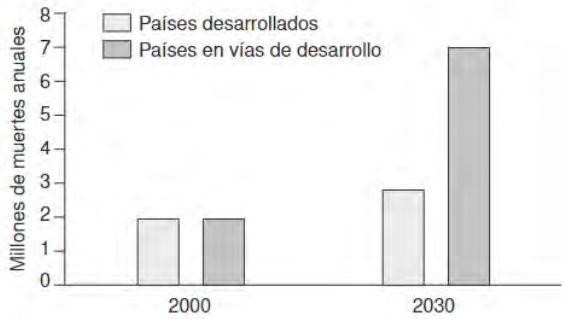

Gráfico 8 Mortalidad según países en desarrollo y en vías de desarrollo

Fuente: Facultad de Medicina Universidad de Cantabria

Es importante mencionar el factor de la edad a la cual los fumadores fallecen, esto está muy relacionado a lo que se expone en el gráfico 05, existen personas que fallecen a la edad de 85 años por algún problema relacionado al consumo de tabaco, pero el problema radica en que alrededor del 40% de las muertes ocurren entre los 35 y 65 años; es decir, en plena edad productiva. Esto conlleva a que la persona no pueda aportar económicamente, aunque el tabaquismo ocasiona muchas muertes súbitas, estas personas igualmente disminuyen su calidad de vida lo cual afecta su productividad y puede traer más gastos para el estado. Relacionando estos datos con el gráfico 05 se puede decir que en la juventud suele darse un mayor consumo de cigarrillos lo cual disminuye su esperanza de vida significativamente en el futuro.

Otro factor importante para considerar es el humo del tabaco que se dispersa en el aire, ya que es responsable de causar una serie de enfermedades tanto en niños como en adultos. El humo del tabaco se considera el principal contaminante ambiental en muchas ciudades, afectando gravemente la salud pública.

Este humo al que muchas personas está expuesta puede conllevar a una pérdida de años de vida, como lo expuesto en el gráfico 09 donde se relaciona el número

de cigarrillos, y los años de vida perdidos entre fumadores y no fumadores, en donde se observa disminución de años de vida mayor en no fumadores que en fumadores.

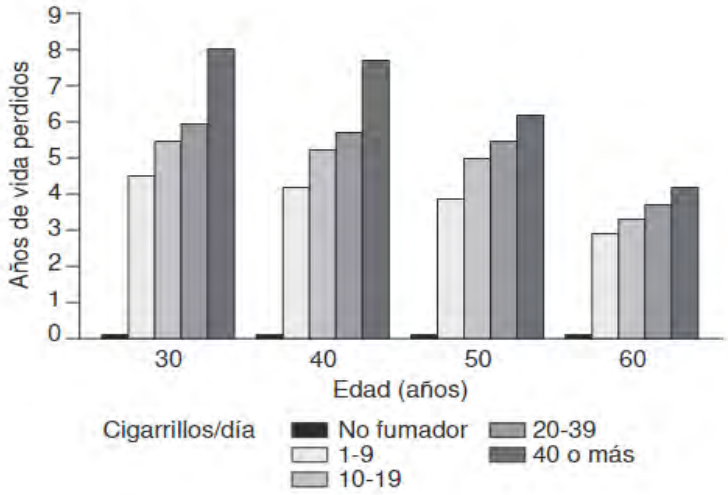

Gráfico 9 Años de vida perdidos en fumadores varones

Fuente: Facultad de Medicina Universidad de Cantabria

Aunque muchos fumadores temen al cáncer de pulmón, este no es el principal responsable de las muertes atribuibles al tabaquismo. En realidad, las enfermedades cardiovasculares son las que causan la mayor cantidad de muertes relacionadas con el consumo de tabaco. En el gráfico 10 se puede observar el impacto del tabaquismo en el aumento del riesgo de desarrollar enfermedades cardiovasculares. Asimismo, en el gráfico 11 se muestra que el riesgo de una persona de desarrollar una enfermedad coronaria es significativamente mayor si es fumadora, siendo esta la principal causa de muerte.

Podemos asignar los siguientes porcentajes a las muertes causadas por el consumo de tabaco: 40% por enfermedades cardiovasculares, 20% por cáncer de pulmón, 5% por otros tipos de cáncer, 25% por enfermedades pulmonares y 10% por otras causas. Con esta información, tenemos como resultado que las enfermedades cardiovasculares son el principal motivo de muerte relacionada con el tabaquismo.

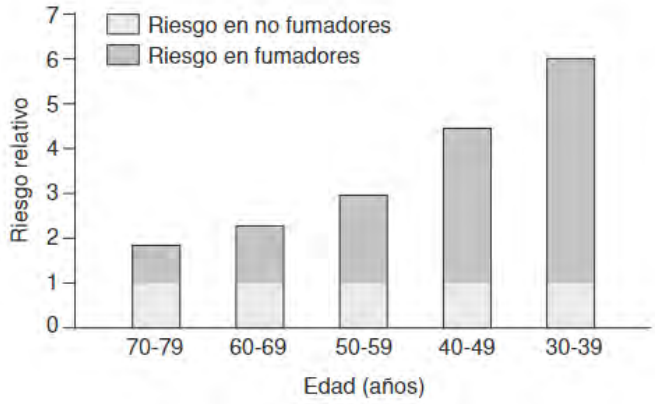

Gráfico 10 Riesgo relativo de accidente coronarios

Fuente: Facultad de Medicina Universidad de Cantabria

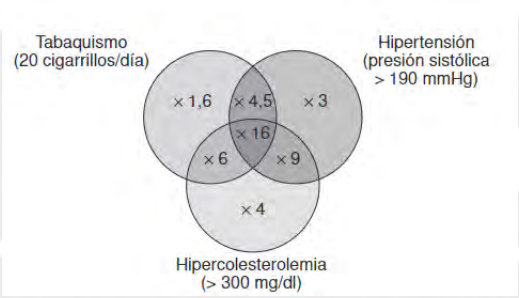

Gráfico 11 Esquema de Kannel potenciación

enfermedades cardiovasculares

Fuente: Facultad de Medicina Universidad de Cantabria

Con lo mencionado anteriormente, es crucial destacar la relación entre la cantidad de cigarrillos que una persona consume y la tasa de mortalidad asociada, como se muestra en el gráfico 12. Cada cigarrillo no solo reduce el tiempo de vida una persona, sino también la calidad de vida de la misma. A menudo, los efectos no se manifiestan de inmediato, y el consumo continuo genera la necesidad de fumar más para satisfacer la sensación de placer que produce la nicotina. Esto conduce a un aumento significativo en la mortalidad, directamente relacionado con la cantidad de cigarrillos que una persona consume. (CORTIJO & FUENTES-PILA, 2004)

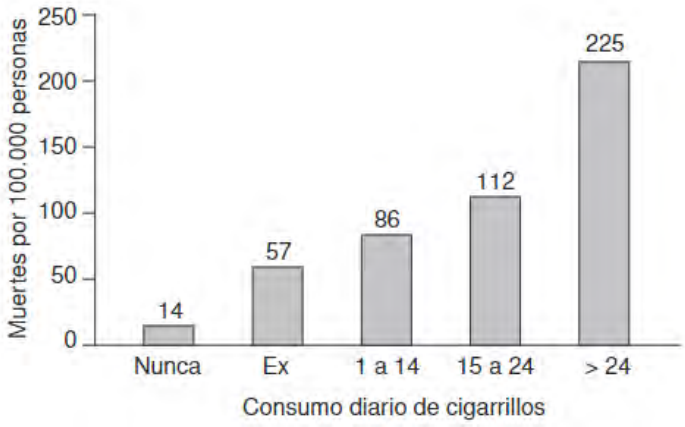

Gráfico 12 Mortalidad por consumo diario de cigarrillos

Fuente: Facultad de Medicina Universidad de Cantabria

2.3 Impacto económico en el Perú debido a los fumadores

El tabaquismo tiene un impacto económico considerable en Perú, abarcando desde la atención médica hasta la pérdida de productividad laboral, así como los gastos en servicios de salud provistos por el estado, entre otros costos que se detallarán más adelante. Según un estudio realizado por la Comisión Nacional Permanente de la Lucha Antitabáquica (COLAT), los costos directos e indirectos del tabaquismo en Perú superan los 15,000 millones de dólares al año. Estos costos incluyen la atención médica para tratar las diversas enfermedades causadas por el tabaquismo, la pérdida de productividad laboral y otros gastos indirectos.

El tabaquismo también afecta negativamente a la economía a largo plazo, ya que la esperanza de vida de los fumadores es más corta, lo que les permite contribuir menos al sistema económico a lo largo de su vida. Además, los fumadores tienen un mayor riesgo de desarrollar discapacidades y enfermedades crónicas, lo que puede afectar sus capacidades laborales y aumentar los costos que el estado debe asumir para tratar estos problemas.

En el gráfico 13, se destaca que el mayor consumo de tabaco proviene de las mujeres, quienes también presentan una menor tasa de abandono del hábito. Fumar también puede afectar a las mujeres gestantes, lo que genera otros costos adicionales para el sistema de salud y reduce la esperanza de vida de la población.

(Moreno, 2017)

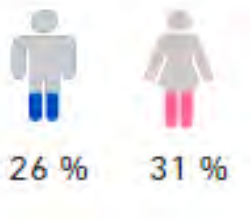

Gráfico 13 Consumo de tabaco Hombres y mujeres

Fuente: Diario Gestión

Además, según un estudio realizado por el COLAT se tiene la tabla 03:

Tabla 3 Consumo de tabaco Hombres y mujeres

%	Población peruana
27%	Es fumadora
40%	Se expone involuntariamente al humo

Fuente: Diario Gestión

Existe un alto porcentaje de personas que, debido al desconocimiento, padecen enfermedades que impactan económicamente en su estilo de vida. Además, los fumadores regulares suelen estar más propensos a desarrollar estas enfermedades (COLAT).

El impacto del consumo de tabaco es directamente responsable de la pérdida de 396,069 años de vida cada año y explica el 12.5% de todos los decesos registrados en el país en personas mayores de 35 años. Esto equivale a 16,719 fallecimientos anuales que podrían haberse evitado.

El costo directo asociado al tabaquismo en Perú asciende a 2,500 millones de dólares anuales, lo que representa el 4% del Producto Bruto Interno (PBI) del país, y el 7.8% del total del gasto en salud cada año.

En la tabla 04 se puede observar que la recaudación por la venta de cigarrillos en el país equivale a 231 millones de dólares anuales, una cifra que apenas cubre poco más del 9% de los gastos provocados por el tabaquismo en el sistema sanitario nacional. (Radovic, 2019)

Tabla 4 Costos e impuestos tabaquismo (en millones de soles)

Costos	Impuestos	% cubre con comisiones
2500	231	9.24%

Fuente: Elaboración propia

La Comisión Nacional Permanente de Lucha Antitabáquica (COLAT) también señala que la cantidad de impuestos recaudados por la industria tabaquera cubre apenas una mínima parte de los costos asociados a los daños causados por el consumo de tabaco, como se muestra en la tabla 03. COLAT menciona que en Perú no existe una regulación efectiva en la publicidad del tabaco ni se han implementado medidas significativas para reducir su consumo. Además, el país carece de estrategias o planes a futuro para abordar este problema.

Al profundizar más en el tema, se observa en el gráfico 14 que en Perú la enfermedad que genera los mayores gastos relacionados con el tabaquismo es la enfermedad pulmonar obstructiva crónica (EPOC). Esta condición provoca una reducción en el flujo de aire y serios problemas respiratorios.

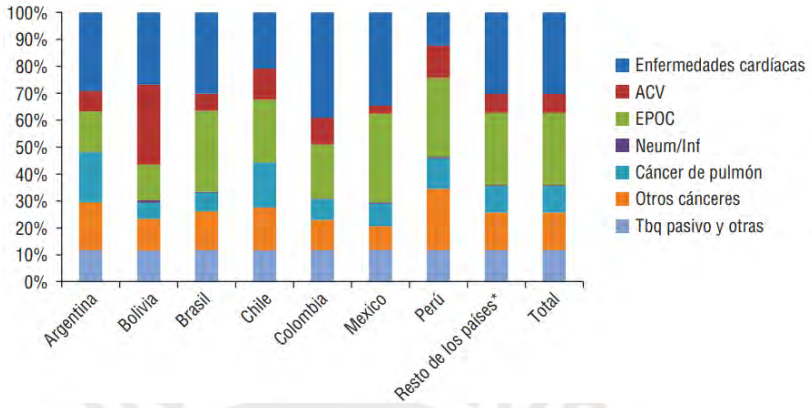

Gráfico 14 Distribución proporcional del gasto atribuible al tabaquismo por causa y país

Fuente: Pan american journey health

Como hemos visto en datos anteriores, el consumo de tabaco tiene efectos económicos. Los estudios han evaluado los impactos económicos del consumo de tabaco en la sociedad, entretanto que otros han destacado la carga económica que los fumadores imponen a los no fumadores. (Pichon-Riviere, Bardach, & Augustovski, 2016)

2.3.1 Enfoques de los costos por el consumo de tabaco

Los costos basados en las prevalencias describen cómo el uso de tabaco en el pasado afecta los costos e ingresos actuales. Por otra parte, los costos basados en las incidencias se centran en los impactos de los consumos actuales de tabaco en los costos futuros, considerando costos e ingresos.

Es importante diferenciar entre los estudios que evalúan el tabaquismo en Perú y los impactos económicos del consumo de tabaco en la sociedad.

Pocas investigaciones se han centrado en el impacto económico de los fumadores sobre los no fumadores o en las consecuencias económicas del uso de los servicios de salud.

Para evaluar los costos externos que el consumo de tabaco ocasiona en la sociedad, como su impacto en las finanzas públicas, es necesario considerar tanto los costos como los ahorros futuros, lo que implica un enfoque basado en las conexiones.

De esta manera, los costos directos incluyen todo lo relacionado con el cuidado de la salud: el gasto realizado por el individuo, la seguridad social o el estado para tratar las enfermedades relacionadas con el tabaco. En cuanto a los costos indirectos, se refiere al impacto social, como la pérdida de productividad debido a enfermedades (pérdida de la capacidad laboral) y una menor esperanza de vida, lo que resulta en una menor contribución a la sociedad.

2.4 Análisis de la problemática

El problema del consumo de tabaco y su detección temprana resulta relevante pues en primer lugar tenemos la alta tasa de mortalidad que lleva a la cifra de que cada año en el Perú 22000 peruanos esto llega a ser cifras preocupantes, además de que en términos económicos, los gastos en el sector salud representan un 4% del PBI, Luego de haber analizado las diferentes problemáticas en el capítulo anterior y haber mencionado las principales podemos ver los beneficios que obtendremos al lograr aplicar medidas preventivas al detectar fumadores en la tabla 05 podemos ver todos los beneficios.

Tabla 5 Beneficios de la detección temprana de fumadores

Beneficios de la detención temprana de los fumadores	
Reducir la mortalidad	Reducir la mortalidad influye directamente en la economía pues la alta tasa de muertes a temprana

	edad se reduciría, reduciendo la cantidad en aproximadamente un 30%.
Reducir el gasto en salud	Al detectar temprano a los fumadores y ayudarlos a dejar el tabaco, podemos reducir el alto costo que conlleva en el sector salud, el cual representa 2500 millones de soles y solo se logra cubrir menos del 10%.
Reducción a la exposición al humo de segunda mano	Al reducir la cantidad de fumadores, los fumadores pasivos también reducirían, los cuales en muchas ocasiones resultan más afectados que los propios fumadores.
Calidad del aire mejorado	El humo del cigarrillo contamina el aire, lo cual produce más smog en el ambiente, reducir los fumadores ayudará a tener más calidad del aire.

Fuente: Elaboración propia

2.5 Diagnóstico

2.5.1 ¿Por qué es relevante?

Actualmente según la OMS sólo el 4% de los fumadores que tienen la intención de dejar de fumar lo logra, algunos de los factores relevantes se muestran en la tabla 06:

Tabla 6 Factores más importantes que dificultan el dejar de fumar

Factores más importantes que dificultan el dejar de fumar	
1	Desconocimiento de los diferentes peligros que conlleva fumar
2	La alta adicción de la nicotina y sus diferentes efectos en el organismo
3	El desconocimiento de estar siendo afectado pasivamente por un medio externo
4	La sociedad en la que el individuo se desarrolla

Fuente: Elaboración propia

Basándonos en estos factores podemos prevenir el desconocimiento de manera oportuna si logramos detectar a tiempo, además de poder detectar a personas que están siendo afectadas y desconocen este hecho. (World Health organization, 2022)

2.5.2 Causas del que problema exista

La libertad que se le da a las empresas distribuidoras de cigarrillos es muy alta, donde no se controla la cantidad de publicidad relacionada, muchas veces enfocada en traer a personas jóvenes, la cantidad de impuestos que paga estas empresas resulta en una perdida pues el gasto que demanda resulta preocupantemente mayor que no logra cubrir ni el 10% del gasto.

El alto nivel de adicción de la nicotina a largo plazo, después de comenzar a consumir cigarrillos la persona siente la necesidad de seguir consumiendo en mayores cantidades para experimentar en la misma intensidad las sensaciones que conlleva su consumó.

El desconocimiento sobre las enfermedades que puede generar comenzar a fumar, muchas personas no conocen muchas de las enfermedades que conlleva fumar, ni todos los problemas que tendrán en el futuro.

Cuando un individuo es afecto pasivamente y no lo detecta a tiempo, muchas veces estamos expuestos al humo del cigarrillo y corremos un mayor riesgo de contraer enfermedades relacionadas.

2.5.3 Diagrama causa-efecto

Se realizará un diagrama causa efecto (Gráfico 15) sobre las causas que impiden que un fumador evite dejar de fumar o sienta la necesidad de empezar a consumir tabaco.

Gráfico 15 Diagrama causa efecto

Fuente: Elaboración propia

2.5.4 Diagrama de Pareto

Luego de analizar las diferentes causas, se realiza un análisis mediante Pareto puntuando cada una de las causas, visualizándose en el gráfico 16:

Causas
Desconocimiento de enfermedades relacionadas
Desconocimiento por parte de fumadores pasivos
Nicotina altamente aditiva
Exceso de publicidad relacionada el tabaquismo
Poca regulación de las empresas productoras de tabaco
Sociedad normalizando el habito
Grupos sociales promoviendo el consumó

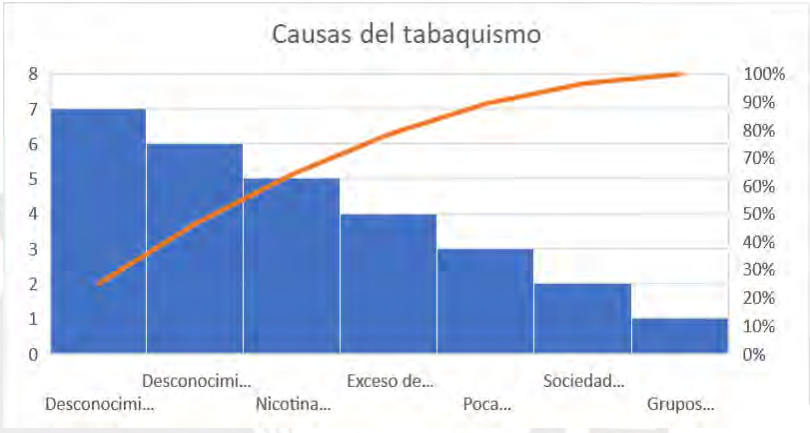

Gráfico 16 Diagrama de Pareto

Fuente: Elaboración propia

Analizando las causas principales en el gráfico 16, se obtiene que el desconocimiento resulta un factor muy importante para lograr combatir el consumo de tabaco, por ello la detección temprano seguida de un plan informativo ayudaría a combatir el tabaquismo, además de la detección de posibles personas que se encuentren expuestas y no lo sepan.

Capítulo 3. Procesamiento y análisis de datos

Como se ha demostrado en los anteriores capítulos fumar afecta negativamente en varios aspectos, por ello se ha propuesto y promovido el tratamiento basado en evidencia para ayudar a dejar de fumar. Sin embargo, solo un tercio de los participaste logro la meta de abstinencia. Muchos médicos encontraron que el asesoramiento para dejar de fumar era ineficaz y lento, y no lo hacía de forma rutinaria en la práctica diaria. Por este motivo, para superar este problema se propuesto varios factores para poder identificar los fumadores que tienen más posibilidades de dejar de fumar. De esta manera para lograr el objetivo de reducir la cantidad de personas que fuman se analizaran los bioseñales para de esta manera identificar el tabaquismo de estas personas.

3.1 Tabla de datos

En la siguiente tabla 07 se da una descripción de las variables que se utilizaran posteriormente en los diferentes modelos.

Tabla 7 Tabla de datos

		Descripción
N	ID	Identificación o índice.
C	gender	Género.
N	age	Edad del paciente
N	height(cm)	Altura en centímetros.
N	weight(kg)	Peso en kilogramos.
N	waist(cm)	Circunferencia de la cintura en centímetros.
N	eyesight(left)	Agudeza visual del ojo izquierdo.
N	eyesight(right)	Agudeza visual del ojo derecho.
C	hearing(left)	Capacidad auditiva del oído izquierdo.
C	hearing(right)	Capacidad auditiva del oído derecho.
N	systolic	La presión sistólica es cuando los ventrículos bombean sangre fuera del corazoón
N	relaxation	Presión arterial diastólica.
N	fasting blood sugar	Nivel de azúcar en sangre en ayunas.
N	Cholesterol	Colesterol total en sangre.
N	triglyceride	Triglicéridos en sangre.
N	HDL	Colesterol tipo HDL.
N	LDL	Colesterol tipo LDL.
N	hemoglobin	Hemoglobina en sangre.
N	Urine protein	Proteína en la orina.
N	serum creatinine	Creatinina sérica en sangre.
N	AST	Enzima AST (glutamic oxaloacetic transaminasa).Enzima que se encuentra principalmente en el hígado
N	ALT	Enzima ALT (glutamic oxaloacetic transaminasa).Enzima que se encuentra en el hígado
N	Gtp	Gama-GTP (gamma-glutamil transferasa).Enzima que se encuentra en el hígado
C	oral	Estado del examen oral.
C	dental caries	Caries dental.
C	tartar	Estado del sarro dental.
C	smoking	Hábito de fumar.

Fuente: Elaboración propia

En la siguiente parte se darán las diferentes características de las variables antes descritas.

3.2 Características de los datos

Se mostrará las diferentes características de las variables en la tabla 08 entre estas tenemos el mínimo, el máximo, la moda, la media, el promedio, los valores faltantes y finalmente el tipo de variable.

Tabla 8 Tabla de datos

la de

d

Name	Mean	Mode	Median	Dispersion	Min.	Max.	Missing	Tipo de variable
N ID	27845.5	0	27845.5	0.58	0	55691	0 (0 %)	Continua
C gender		M		0.656			0 (0 %)	Binaria
N age	44.18	40	40	0.27	20	85	0 (0 %)	Continua
N height(cm)	164.65	170	165	0.06	130	190	0 (0 %)	Continua
N weight(kg)	65.86	65	65	0.19	30	135	0 (0 %)	Continua
N waist(cm)	82.0464	80	82	0.113035	51	129	0 (0 %)	Continua
N eyesight(left)	1.01262	1.2	1	0.4808	0.1	9.9	0 (0 %)	Continua
N eyesight(right)	1.00744	1.2	1	0.48237	0.1	9.9	0 (0 %)	Continua
C hearing(left)		1		0.119			0 (0 %)	Binaria
C hearing(right)		1		0.121			0 (0 %)	Binaria
N systolic	121.494	110	120	0.112564	71	240	0 (0 %)	Continua
N relaxation	76.0048	80	76	0.12735	40	146	0 (0 %)	Continua
N fasting blood sugar	99.3123	94	96	0.209394	46	505	0 (0 %)	Continua
N Cholesterol	196.901	199	195	0.184344	55	445	0 (0 %)	Continua
N triglyceride	126.666	71	108	0.565577	8	999	0 (0 %)	Continua
N HDL	57.2903	54	55	0.257266	4	618	0 (0 %)	Continua
N LDL	114.965	110	113	0.355989	1	1860	0 (0 %)	Continua
N hemoglobin	14.6226	15	14.8	0.106991	4.9	21.1	0 (0 %)	Continua
N Urine protein	1.08721	1	1	0.372401	1	6	0 (0 %)	Continua
N serum creatinine	0.885738	0.9	0.9	0.250099	0.1	11.6	0 (0 %)	Continua
N AST	26.1829	20	23	0.739233	6	1311	0 (0 %)	Continua
N ALT	27.036	15	21	1.14468	1	2914	0 (0 %)	Continua
N Gtp	39.9522	15	25	1.25876	1	999	0 (0 %)	Continua
C oral		Y		0			0 (0 %)	Constante
C dental caries		0		0.518			0 (0 %)	Binaria
C tartar		Y		0.687			0 (0 %)	Binaria
C smoking		0		0.657			0 (0 %)	Binaria

Fuente: Elaboración propia

En la siguiente parte se realizará la construcción de los datos donde se realizará las diferentes modificaciones a la base de datos para evaluarla con los modelos posteriormente.

3.3 Construcción de los datos

Para la construcción de los datos se procede a construir la base eliminamos las variabas que identificamos no resultan significativas en este caos la variable ´´Oral´' la cual tiene la distribución en el grafico 17 el valor es el mismo para todos, debido a que el único significado de esta variable es verificar si es que se ha realizado una revisión bucal, en este caso todos han realizado una revisión bucal.

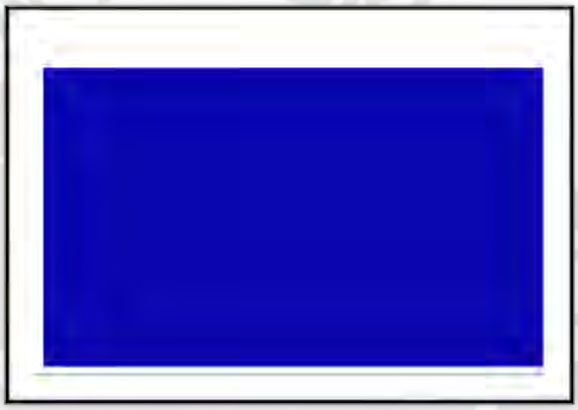

Gráfico 17 Distribución de la variable Oral

Fuente: Elaboración propia

Además, se calcula la variable índice de masa corporal (IMC), una medida que se utiliza para evaluar si una persona tiene un peso saludable en relación con su altura. La cual se calcula con la fórmula mostrado en el gráfico 18. Para esto se utiliza el peso y la altura. Por ello se eliminarán estas variables pues resultan redundantes debido al nuevo cálculo.

$$IMC = \frac{PESO\ (KILOGRAMOS)}{(ALTURA\ EN\ METROS)^2}$$

Gráfico 18 Formula IMC

Fuente: Elaboración propia

3.4 Correlación entre la variable

Se procede a calcular la correlación de las variables identificando las variables binarias y las continuas.

3.4.1 Matriz de correlación de pearson

Este coeficiente mide la relación lineal entre pares de variables continuas, indicando si hay una asociación positiva, negativa o nula entre ellas, en el gráfico x se identifica que existe una alta relación entre algunas variables (waist-IMC),(LDL-Cholesterol),(syntolic-relaxation) y (ALT-AST). Por ello se eliminará Waist, LDL, Syntolic y ALT, para evitar que los modelos resulten alterados por estos valores, esto se observa en el gráfico 19.

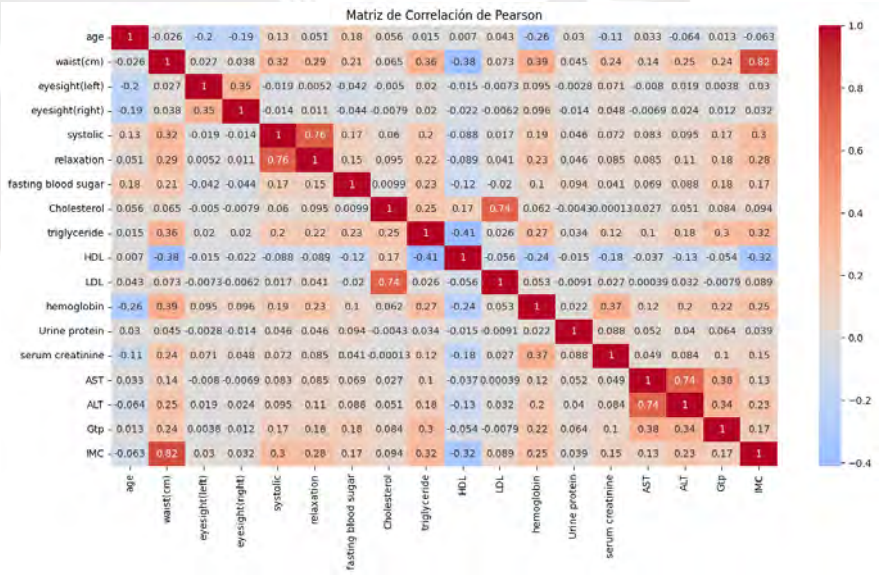

Gráfico 19 Matriz de correlación de pearson

Fuente: Elaboración propia

En la siguiente parte se realizará la correlación con las variables binarias.

3.4.2 Matriz de correlación de punto biseral

La matriz de correlación de punto biserial es una matriz que muestra las correlaciones entre una variable binaria (que solo tiene dos categorías) y una variable continua. Como se observa en el gráfico 20. las variables que poseen una alta relación entre hemoglobina y género, por ello se eliminará la variable hemoglobina.

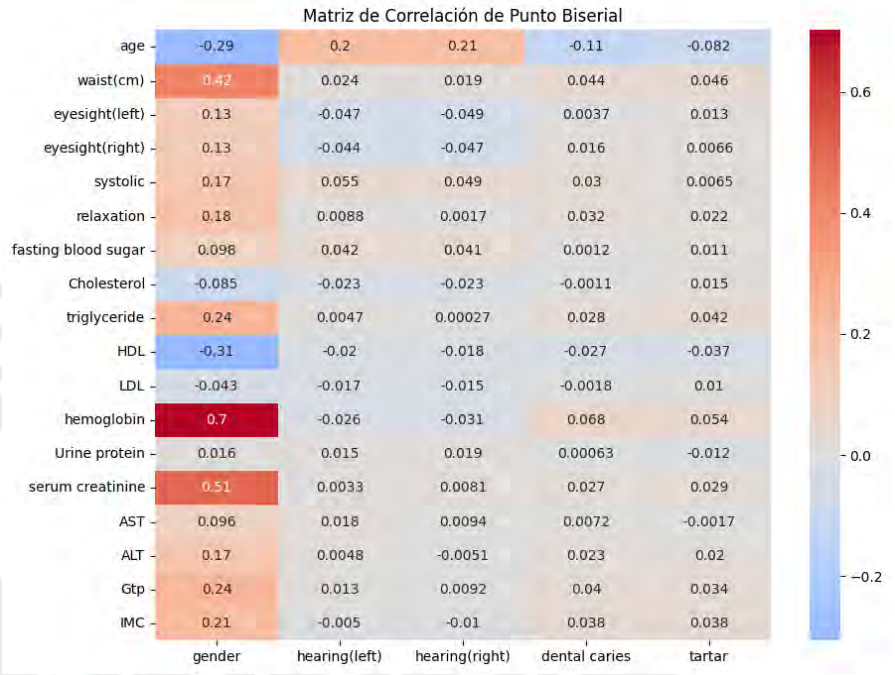

Gráfico 20 Matriz de correlación de punto biseral

Fuente: Elaboración propia

Luego de eliminar todas las variables tenemos en total, quedando en total 20 variables que serán evaluadas en los siguientes modelos, obteniendo los siguientes resultados de precisión en la tabla 09.

Tabla 9 Precisión entre modelos

	Precisión	
Modelos	Eliminando variables	Sin eliminar variables
Arboles de decisión	82.774%	83.243%
Random Forest	85.299%	85.845%
Regresión logística	72.196%	75.751%
Máquina de vectores de soporte	81.351%	81.631%
Extreme Gradient Boosting	83.309%	83.581%
Redes neuronales	84.261%	83.881%

Fuente: Elaboración propia

Por ello se utilizará las variables sin eliminar para todos los modelos, menos para el modelo de redes neuronales donde se eliminarán las variables antes mencionadas.

Capítulo 4. Propuesta de modelo predictivo

En la siguiente sección se procederá a evaluar y construir los modelos anteriormente descritos.

4.1 Selección de la técnica de modelado

Se evaluará los diferentes modelos, para encontrar los mejores parámetros para cada modelo.

4.1.1 Arboles de decisión

Los Árboles de Decisión son una técnica ampliamente utilizada en la modelización de problemas de clasificación. Esta técnica se basa en la construcción de un árbol que toma decisiones basadas en reglas. La calidad de las divisiones en el árbol se mide utilizando diferentes criterios, uno de los cuales es la entropía, que se emplea para calcular la impureza de los nodos y seleccionar las divisiones óptimas.

Después de implementar el código correspondiente, obtuvimos un árbol de decisiones con múltiples opciones debido a la cantidad de variables y escenarios. Al representar gráficamente solo las primeras divisiones, se obtuvo el resultado mostrado en el gráfico 21. En este gráfico, se observa que el nodo raíz se basa en el género, donde se produce una división según el género. En este caso, se definió el género con el valor de 1 para masculino

y 0 para femenino. Si el valor es menor o igual a 0.5, es decir, si es femenino, la ruta a seguir es hacia la derecha, mientras que si es masculino, se toma el camino hacia la izquierda. Posteriormente, se presentan las demás variables en función de los valores obtenidos en las diferentes ramificaciones del árbol.

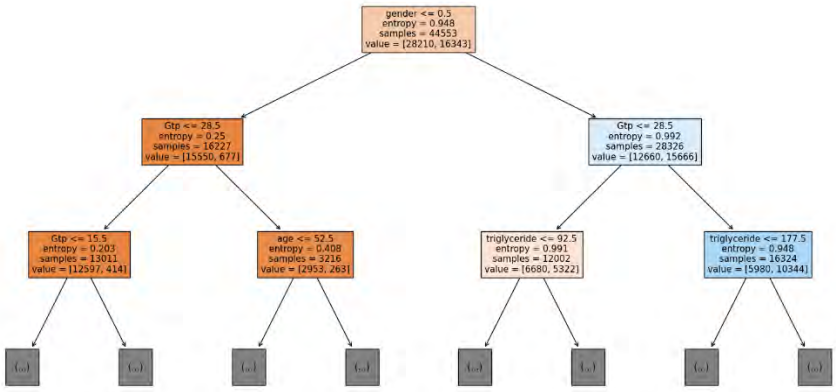

Gráfico 21 Árbol de decisión

Fuente: Elaboración propia

También generamos la matriz de confusión correspondiente, la cual se utilizará posteriormente para evaluar el mejor modelo, como se muestra en la Tabla 10.

Tabla 10 Matriz de confusión

	Sí	No
Sí	5827	1200
No	1173	2939

Fuente: Elaboración propia

El siguiente modelo que se evaluará es el de Random Forest. Posteriormente, se llevará a cabo una comparación entre los modelos.

4.1.2 Random Forest

Random Forest es un poderoso algoritmo de aprendizaje automático utilizado tanto para la clasificación como para la regresión. Este método se basa en la construcción de múltiples árboles de decisión y la combinación de sus predicciones para mejorar la precisión y reducir el sobreajuste.

Al implementar este algoritmo, se vuelve algo más complejo que un Árbol de Decisión, ya que es necesario determinar los mejores parámetros para el modelo. Uno de los aspectos clave es encontrar el número óptimo de árboles de decisión. Para ello, se evalúa una lista que contiene diferentes valores para la cantidad de árboles, como se muestra en la tabla 11.

Tabla 11 Valores evaluados

Valores evaluados	10	50	100
	150	200	

Fuente: Elaboración propia

De la tabla 12, mediante una validación cruzada se obtienen puntuaciones para la lista evaluada, y se concluye que el resultado óptimo es el de 200 árboles de decisión.

Tabla 12 Valores evaluados

Valores evaluados	10	50	100	150	200
Puntuación	0.706	0.7315	0.7342	0.7316	0.7357

Fuente: Elaboración propia

Luego de definir los mejores parámetros se procede a construir el modelo y sacamos la matriz de confusión correspondiente en la tabla 13.

Tabla 13 Matriz de confusión

	Sí	No
Sí	5798	1229
No	956	3156

Fuente: Elaboración propia

En la siguiente parte se evaluará los datos con el modelo de regresión logística.

4.1.3 Regresión logística

La Regresión Logística es una técnica fundamental en estadística y aprendizaje automático utilizada para resolver problemas de clasificación. A diferencia de los modelos de regresión lineal que predicen valores numéricos, la Regresión Logística se utiliza para predecir probabilidades de pertenencia a una o más categorías discretas.

En primer lugar, se realizará un gráfico de dispersión, como se puede observar en el Gráfico 22, para analizar el comportamiento de las variables. A través de este gráfico, podemos identificar si existe alguna relación entre estas variables y determinar si se pueden identificar grupos representativos en los datos.

Gráfico 22 Gráfico de dispersión

Fuente: Elaboración propia

Del gráfico, parece haber una relación entre las variables, lo que sugiere que la Regresión Logística podría ser un buen método para modelar estos datos. Como sabemos, el modelo evaluará las variables independientes de entrada y utilizará la función sigmoide para transformar estas variables y asignarles un valor entre 0 y 1. En este caso, se utiliza el valor 1 para representar a pacientes fumadores y 0 para representar a pacientes no fumadores.

Después de construir el modelo de Regresión Logística, se obtiene la matriz de confusión correspondiente, como se muestra en la tabla 14. Esta matriz de confusión es fundamental para evaluar el rendimiento del modelo y entender cómo se comporta en términos de clasificación.

Tabla 14 Matriz de confusión

	Sí	No
Sí	5701	1326
No	1825	2287

Fuente: Elaboración propia

Ahora procederemos a realizar la evaluación del modelo mediante un modelo de máquinas de vectores de soporte.

4.1.4 Máquina de vectores de soporte

Las Máquinas de Vectores de Soporte (SVM) son un potente método de aprendizaje supervisado utilizado en problemas de clasificación y regresión. Su objetivo principal es identificar el hiperplano que mejor separa dos clases de datos en un espacio multidimensional, maximizando la distancia entre los puntos de datos más cercanos a cada clase.

La selección de los parámetros óptimos es fundamental al utilizar este modelo. Para lograrlo, se evalúa la precisión utilizando diferentes kernels, que son funciones matemáticas que miden la similitud o distancia entre dos puntos en un espacio de dimensiones superiores. En este caso, se evaluarán los siguientes kernels: Kernel Lineal, Kernel Polinomial, Kernel de Función de Base Radial (RBF) y Kernel Sigmoide.

Para cada uno de estos kernels, aplicaremos el algoritmo SVM a nuestros datos y obtendremos la matriz de confusión y la precisión de cada modelo. Los resultados se presentan en las siguientes tablas:

- Kernel Lineal en la Tabla 15.
- Kernel Polinomial en la Tabla 16.
- Kernel de Función de Base Radial en la Tabla 17.
- Kernel Sigmoide en la Tabla 18.

De acuerdo con los resultados, se observa que la mejor puntuación se obtiene con el Kernel Lineal, lo que sugiere que este kernel es el más adecuado para el conjunto de datos y la tarea de clasificación.

Tabla 15 Matriz de confusión (Kernel Lineal)

	Sí	No
Sí	5346	1681
No	1203	2909

Presicion	81.631%

Fuente: Elaboración propia

Tabla 16 Matriz de confusión (Kernel Polinomial)

	Sí	No
Sí	5975	1052
No	2016	2096

Presicion	74.772%

Fuente: Elaboración propia

Tabla 17 Matriz de confusión (Kernel Radial)

	Sí	No
Sí	5997	1030
No	2124	1988

Presicion	73.846%

Fuente: Elaboración propia

Tabla 18 Matriz de confusión (Kernel Sigmoide)

	Sí	No
Sí	4442	2585
No	2676	1436

Presicion	62.405%

Fuente: Elaboración propia

El siguiente modelo para evaluar será el Extrteme Gradient Boosting.

4.1.5 Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) es un potente algoritmo de aprendizaje automático que se utiliza tanto en problemas de clasificación como de regresión. Su principal fortaleza radica en su capacidad para construir conjuntos de árboles de decisión de manera secuencial, corrigiendo los errores de predicción de los árboles anteriores.

El rendimiento de este algoritmo depende en gran medida de la correcta selección de los parámetros, que incluyen:

- n_estimators: Número de árboles (estimadores) en el modelo.
- max_depth: Profundidad máxima de cada árbol en el modelo.

- learning_rate: Tasa de aprendizaje, que controla la contribución de cada árbol al modelo final.
- subsample: Fracción de muestras utilizadas para entrenar cada árbol.
- colsample_bytree: Fracción de características (columnas) utilizadas para entrenar cada árbol.

Para encontrar los mejores hiperparámetros, utilizamos las siguientes combinaciones y aplicamos el estimador GridSearchCV, que realiza la búsqueda de hiperparámetros de manera exhaustiva. Los mejores parámetros encontrados son los siguientes:
Mejores hiperparámetros encontrados:

- 'colsample_bytree': 1.0
- 'learning_rate': 0.2
- 'max_depth': 5
- 'n_estimators': 300
- 'subsample': 0.8

Posteriormente, definimos estos parámetros óptimos para ejecutar el algoritmo, lo que resulta en la matriz de confusión que se muestra en la tabla 19. Esta matriz de confusión es esencial para evaluar el rendimiento del modelo y comprender cómo se comporta en términos de clasificación.

Tabla 19 Matriz de confusión

	Sí	No
Sí	5728	1299
No	1139	2973

Fuente: Elaboración propia

Finalmente evaluaremos nuestros datos con el modelo de redes neuronales.

4.1.6 Redes neuronales

Las redes neuronales son modelos de aprendizaje automático inspirados en las funciones del cerebro humano. Consisten en capas interconectadas de

"neuronas" artificiales que procesan información y aprenden patrones a partir de datos.

Las redes neuronales tienen una serie de parámetros que pueden ajustarse para optimizar su rendimiento. Estos parámetros suelen ser más complejos que los de los modelos evaluados anteriormente. Para facilitar la comparación, denominaremos al primer modelo "Modelo A" y al segundo modelo "Modelo B".

En el Modelo A, utilizamos GridSearchCV para obtener los mejores hiperparámetros, que son los siguientes:

- Hidden_layer_sizes: Este parámetro determina la arquitectura de la red, es decir, cuántas capas ocultas tiene el MLP y cuántas neuronas hay en cada capa. Se prueba con tres opciones: (128, 64), (256, 128) y (128,). Estos valores representan diferentes configuraciones de capas y neuronas ocultas.
- Activation: Este parámetro especifica la función de activación que se aplica a la salida de cada neurona en las capas ocultas. Se prueban dos opciones: 'relu' (Rectified Linear Unit) y 'tanh' (tangente hiperbólica).
- Alpha: Este hiperparámetro controla la cantidad de regularización aplicada a los pesos de la red para evitar el sobreajuste. Se prueban tres valores: 0.0001, 0.001 y 0.01.

Luego el valor óptimo de los hiperparámetros es el siguiente:

- Hidden_layer_sizes: 256, 128
- Activation: tanh
- Alpha: 0.0001

Luego los valores obtenidos en la matriz de confusión y su precisión correspondiente se observan en la tabla 20.

Tabla 20 Matriz de confusión (modelo A)

	Sí	No
Sí	5921	1106
No	1106	3006

Presición	84.261%

Fuente: Elaboración propia

Ahora, al evaluar el Modelo B, utilizamos KerasClassifier para la búsqueda de hiperparámetros. Los hiperparámetros a optimizar en este caso son los siguientes:

- Hidden_layer_sizes: Este parámetro controla la arquitectura de la red, es decir, cuántas capas ocultas tiene el MLP y cuántas neuronas hay en cada capa. Se probaron tres opciones: (128, 64), (256, 128) y (128,). Estos valores representan diferentes configuraciones de capas y neuronas ocultas.
- Activation: Este parámetro especifica la función de activación que se aplica a la salida de cada neurona en las capas ocultas. Se probaron dos opciones: 'relu' (Rectified Linear Unit) y 'tanh' (tangente hiperbólica).
- Dropout: El parámetro dropout controla la tasa de abandono (dropout rate) que se aplica después de cada capa densa para regularizar el modelo y prevenir el sobreajuste. Se probaron dos valores: 0.2 y 0.3, que representan diferentes tasas de abandono.
- Epochs: Este parámetro determina el número de épocas o iteraciones completadas durante el entrenamiento de la red neuronal. Se probaron tres valores: 10, 20 y 30, que representan diferentes números de épocas de entrenamiento.
- Batch_size: El parámetro batch_size define el tamaño del lote utilizado en cada paso de actualización de pesos durante el entrenamiento. Se probaron dos tamaños de lote: 32 y 64, que representan diferentes tamaños de lote.

  Ahora los mejores paramatros a evaluar obtenidos osn los siguientes:

  Los mejores valores obtenidos son los siguientes:

- Activation: 'tanh'
- Deopout: 0.2
- Hidden_Layer_Sizes: (128,64)
- Batch_Size: 32
- Epochs: 20

Luego del modelo B se obtiene la siguiente matriz de confusión en la tabla 21:

Tabla 21 Matriz de confusión (modelo B)

	Sí	No
Sí	5597	1430
No	1246	2866

Presición	81.792%

Fuente: Elaboración propia

Luego de comparar la precisión entre el Modelo A y el Modelo B, se observa que el Modelo A proporciona una mejor precisión. Por lo tanto, hemos seleccionado el Modelo A como la mejor opción de Redes Neuronales en función de los resultados obtenidos.

En el siguiente capítulo se comparan los diferentes métodos descritos y se describirán el proceso que se seguirá después de detectar a las personas fumadoras.

4.2 Evaluación cuantitativa de modelos predictivos

En esta sección, la matriz de confusión se utiliza para evaluar y calcular la precisión de cada modelo, lo que permitirá comparar y determinar cuál es el mejor.

4.2.1 Evaluación de resultados

Para evaluar los resultados se calculará la precisión cuya fórmula podemos observarla en la tabla 22.

Tabla 22 Fórmula de precisión

		Predecido	
		1	0
Actual	1	TP	FN
	0	FP	TN

Precisión = TP ÷ (TP + FP)

Fuente: Elaboración propia

Luego mediante la predicción se evaluará cada uno de los modelos.

4.2.2 Prueba de precisión

Luego de calcular en la precisión para cada modelo se identifica como mejor modelo Random Forest, con la mayor precisión en la tabla 23.

Tabla 23 Precisión de los modelos

Modelos	Precisión
Arboles de decisión	83.243%
Random Forest	85.845%
Regresión logística	75.751%
Máquina de vectores de soporte	81.631%
Extreme Gradient Boosting	83.581%
Redes neuronales	84.261%

Fuente: Elaboración propia

Los resultados obtenidos serán analizados en el siguiente capítulo y se evaluara toda la información recopilada hasta el momento.

Capítulo 5. Evaluación de resultados

En la siguiente sección se estimará cual sería el impacto obtenido aplicando el modelo predictivo obtenido anteriormente.

5.1 Estimación económicas de la aplicación del modelo

Tomando como referencia que la enfermedad que más aflige a las personas fumadoras es el EPOC y con la información recaudada anteriormente se estimara lo siguiente.

Partiendo de la cantidad de población peruana fumadora y fumadores pasivos, podemos determinar dos grupos los cuales se muestran en la tabla 24. Teniendo en cuenta que actualmente la población peruana es de 32626000 habitantes (OPS, 2023).

Tabla 24 % de fumadores 2023

%	Población peruana	Población
27%	Es fumadora	8809020
40%	Se expone involuntariamente al humo	13050400

Fuente: Elaboración propia

Además, Es importante recordar que alrededor del 10% de la población es diagnosticada con enfermedad pulmonar obstructiva crónica (EPOC), según la Guía peruana de EPOC 2016 como se muestra en la tabla 25, se debe tener en cuenta que con e pasar de los años el número de personas mayores se ha ido incrementando y este porcentaje ha ido creciendo. (Sociedad Peruana de Neumología, 2016)

Tabla 25 % de personas con EPOC

Población actual	33000000
% con EPOC	10%
Población con EPOC	3300000

Fuente: Elaboración propia

Luego de esto se debe considerar los gastos que se tienen debido al EPOC esto abarca desde medicamento, urgencias, oxigeno, entre otros valores. Con ello se llega a la conclusión de que por persona se tienen un gasto de 19985.73 soles en la tabla 26 se puede apreciar cada uno de los gastos más a detalle. (Villarreal, et al.,2018)

Tabla 26 Costo promedio anual en EPOC por tipo de insumo y servicio

Insumo	Costo Promedio				
	Medicina Familiar	Neumología	Hospital	Urgencias	Costo por Insumo
Medicamentos	2855.80	253.29	8186.67	2183.12	13478.88
Oxígeno	4314.39	0.00	102.36	21.75	4438.49
Laboratorio	20.76	59.62	89.62	24.75	194.76
Ecocardiograma	0.00	115.39	9.23	0.00	124.63
Eectrocardiograma	0.00	39.04	11.71	15.62	66.36
Espirometria	0.00	53.43	0.00	0.00	53.43
Radiografía de tórax	14.79	3.59	6.72	8.97	34.07
Tomografía axial computarizada	0.00	0.00	22.12	0.00	22.12
Consultas en medicina familiar	592.30	0.00	0.00	0.00	592.30
Consultas en neumología	0.00	82.38	0.00	0.00	82.38
Hospitalización	0.00	0.00	461.44	0.00	461.44
Interconsulta cardiología	0.00	0.00	4.88	0.00	4.88
Interconsulta a medicina interna	0.00	0.00	5.49	0.00	5.49
Urgencias	0.00	0.00	0.00	126.18	126.18
Costo promedio	7798.04	606.74	8900.23	2380.38	19685.40

Fuente: Rev. Medica del Instituto mexicano del seguro social

Ahora que sabemos que el individual de entender EPOC es de 19985.73 soles equivalente a 5948.13 dólares estadunidenses, debemos ajustar este valor al año actual. Para ello utilizaremos la siguiente formula

$Valor\ actual = Valor\ inicial X(1 + tasa\ de\ inflación)^{Número\ de\ años\ entre\ fechas}$.

Obteniendo los datos de la inflación del Banco Central de Reserva del Perú. Como se muestra en la tabla 27 el costo actual sería de 8461.63 dólares estadunidenses, dándonos un monto con el tipo de cambio de 3.85 soles/dólar de 32577.28 soles en la actualidad.

Tabla 27 Equivalentes de costos al año 2023

Valor inicial dolares	2007	5282.77
Inflación	Oct23	2.988094265
Valor actual	2023	8461.630418
Valor actual soles	cambio: 3.85	32577.27711

Fuente: Elaboración propia

Con esto se tiene aproximadamente un gasto de nueve millones de dólares, lo cual sería un gasto individual que se tendría debido a padecer de EPOC.

5.2 Evaluación del mejor plan de acción

Luego de tener todos estos resultados se tienen dos escenarios uno donde el fumador resulta ser un fumador pasivo, donde el plan a seguir principalmente se basa en informar a la persona que está exponiéndose a grandes cantidades de humo y recomendarle utilizar equipo de protección en caso se exponga en actividades que realiza regularmente o tomar las medidas necesarias para mejorar esta situación.

Por otro lado, se tiene a los fumadores regulares los cuales al ser adictos a la nicotina suelen tener una mayor dependencia a la misma y por ello les cuesta más dejarla. Para ello existen múltiples tratamientos en este caso analizaremos tres tratamientos publicados en el artículo “Combined pharmacotherapy and behavioural interventions for smoking cessation (Review)” en la tabla 28 se presentan los tratamientos con la eficacia que presenta cada uno. (Stead LF, 2016)

Tabla 28 Tratamientos para dejar el tabaco

Tratamiento	Descripción
Farmacológico	Los participantes recibieron bupropion y vareniclina
Conductual	Los participantes recibieron terapia cognitivo-conductual (TCC).
Combinado	Los participantes recibieron bupropion, vareniclina y TCC.

Fuente: Elaboración propia

Este estudio muestra la eficacia del tratamiento combinado para aumentar la probabilidad de que las personas dejen de fumar. entre un 70% a 100%.

Se procederá a calcular los ahorros teniendo en cuenta la aplicación del modelo predictivo.

5.3 Impacto esperado de la mejora

En los siguientes puntos analizará el impacto económico y el efecto en la esperanza de vida que se tendrá aproximadamente.

5.3.1 Impacto económico basado en la enfermedad más común

En la tabla 29 podemos ver la cantidad de personas a las cuales apoyaría el modelo y el ahorro que esto significaría. Donde como se mencionó anteriormente el ahorro individual será de S/ 26,532.40 soles, y en total se ahorra 658482127185.62 soles. Se ha considerado las probabilidades del modelo, como de los tratamientos para el cálculo. Se ha tenido en consideración el porcentaje de personas que realiza regularmente un examen médico cada 12 meses según (ENDES) 2022.

Tabla 29 Estimación de ahorros

Tipo de paciente identificado						
Fumador Activo						
Variables						**Resultado**
Ahorro	Realizan Exámenes médicos	Tratamientos	Mejor modelo	Población peruana	Población	**Ahorro total**
Ahorro individual por persona	% de población que realiza exámenes	% de éxito	Random Forest	% Población fumadora	Cantidad de peruanos actuales	**Producto de todas las variables**
68820.8	58%	70%	86%	27%	32626000	211295623154.64

Fumador pasivo						
Variables						**Resultado**
Ahorro	Realizan Exámenes médicos	Tratamientos	Mejor modelo	Población peruana	Población	**Ahorro**
Ahorro individual por persona	% de población que realiza exámenes	% de éxito	Random Forest	%Fumadores pasivos	Cantidad de peruanos actuales	**Producto de todas las variables**
68820.8	58%	100%	86%	40%	32626000	447186504030.98

Ahorro total:	
Suma de resultados	658482127185.62

Fuente: Elaboración propia

Teniendo en consideración que las personas fumadoras les afectarán el porcentaje del modelo predictivo y el porcentaje de éxito que tenga el tratamiento que le ayudara a dejar de fumar, mientras que por otro lado las personas expuestas al humo de manera indirecta dependerán solo del porcentaje de efectividad del modelo para poder ser detectadas.

5.3.2 Aumento de la esperanza de vida

Para poder determinar cuánto incrementa la esperanza de vida es necesario identificar varios factores, en primer lugar según un estudio publicado por The Lancet en 2013, descubrió que Las personas que dejan de fumar antes de los 40 años tienen un riesgo de muerte 10 años menor en comparación con las personas que continúan fumando.(Zhao & Wang, 2020), mientras que un estudio publicado por la revista JAMA Señala que las personas que dejan de fumar después de los 50 años tienen un riesgo de muerte un 30% menor que quienes continúan fumando. (Blake Thomson, et al., 2022) y finalmente un estudio publicado por la revista American Jornal (American Jornal, 2017) indica que los beneficios de dejar de fumar incrementan a medida va pasando el tiempo, mientras antes de deje de fumar mayor será la esperanza de vida. Esto se puede ver resumido en la tabla 30.

Tabla 30 Esperanza de vida

Edad de dejar de fumar	Aumento de la esperanza de vida
Menos de 40 años	10-15 años
Entre 40 y 50 años	5-10 años
Después de los 50 años	3-5 años

Fuente: Elaboración propia

En la siguiente parte se evaluará las conclusiones y recomendaciones que se obtienen después del análisis.

Capítulo 6. Conclusiones y recomendaciones

En los siguientes puntos se dará pase a las conclusiones y recomendaciones sen base a los estudios realizados.

6.1 Conclusiones

- Implementar este modelo predictivo durante las consultas resultara en una mejor detección de personas que pueden estar exponiéndose al humo o ser fumadores recién iniciando con este hábito debido a que existe un alto porcentaje de estos grupos que pueden beneficiarse con una detección temprana, lo que permite prevenir las enfermedades, mejorar la calidad de vida de la persona y mantener el ciclo económico sin alteraciones.
- Sería recomendable aplicar el flujo presentado en la figura 23, pues los gastos asociados al consumo del tabaco son bastante altos y el ahorro llegaría a ser de casi 7 millones.

Gráfico 23 Flujo según características

Fuente: Elaboración propia

- Se debe evaluar si las regulaciones actuales brindadas por el estado peruano son suficientes, pues lo gastos a raíz del consumo de tabaco no cubren ni el 10% del costo en salud debido a enfermedades por la exposición al humo.

En la siguiente parte se dan algunas recomendaciones a tener en cuenta al momento de aplicar estos nuevos planes.

- Utilizar el modelo puede ayudar a aumentar el porcentaje de detección de EPOC, pues según el estudio de Guía Peruana de EPOC -2016, el 64% de los casos resultan ser diagnósticos erróneos, además de ayudar a identificar otras enfermedades.
- El ahorro basándose solo en la enfermedad más común resulta aproximadamente del 29% como podemos observar en la tabla 31.

Tabla 31 % de ahorro EPOC

Gasto	2245347420800
Ahorro	658482127186
% de ahorro	29%

Fuente: Elaboración propia

6.2 Recomendaciones

- El modelo debe retroalimentarse para que se esta manera mejore y se adecue a los datos de la población local y debe de ajustarse en caso sea necesario.
- El veredicto sobre el tratamiento o recomendaciones la tendrá siempre el doctor y el decidirá cuál es el mejor tratamiento o recomendación.
- Se debe dar un mayor control sobre el tabaco, pues en el Perú este aun cuenta con muchas libertades y la población se encuentra bastante desinformada.
- Se debe realizar estudio para ayudar a las personas que están siendo afectadas indirectamente y de esta manera identificar las principales causas de estos males.

Bibliografía

Alonso, E. D., & Padilla, D. A. (2001). Regresión logística. Un ejemplo de su uso en Endocrinología. *Revista Cubana de Endocrinología*, 7. Obtenido de https://n9.cl/2sij1

American Jornal. (2017). The Health Benefits of Smoking Cessation. *American Jornal*. Obtenido de https://n9.cl/9h72l

Andina. (29 de Mayo de 2019). ¡Alarma! Consumo de tabaco en el Perú afecta principalmente a jóvenes. *Andina*, pág. 1. Obtenido de https://n9.cl/s4ckte

Asencios, V. V. (2004). *DATA MININGYEL DESCUBRIMIENTO DEL CONOCIMIENTO.* Lima: Industrial Data. Obtenido de https://n9.cl/0zw67

Blake Thomson, D., & Emberson, J. (2022). Association Between Smoking, Smoking Cessation, and Mortality by Race, Ethnicity, and Sex Among US Adults. *JAMA*. Obtenido de https://n9.cl/50crcw

Ballesteros, H. F., Iñiguez, E. G., & Velasco, S. R. (2018). *Minería de Datos.* Ecuador: Editorial Saberes del Conocimiento. Obtenido de https://n9.cl/w0afs

Chávez, J. J., Mora, J. D., & Montoya, R. A. (2013). Aplicación del aprendizaje automático con árboles de decisión al estudio de las variables del modelo de indicadores de gestión de las universidades públicas. *Scientia et Technica Año XVIII, Vol. 18, No. 4,*, 7. Obtenido de https://n9.cl/5wwn3

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *University of Washington*, 10. Obtenido de https://n9.cl/oot7n4

CORTIJO, C., & FUENTES-PILA, J. (2004). Toxicidad derivada del consumo de tabacoSmoking-induced toxicity. *Trastornos Adictivos 200*, 89-94. Obtenido de https://n9.cl/md12h

Cutler, A., Cutler, D. R., & Stevens, J. (2014). *Random Forest.* Utah: Utah State University. Obtenido de https://n9.cl/jxkk2c

Enrique Ruiz Mori, H. R.-R.-M.-V.-R. (2016). Conocimiento de los riesgos del tabaquismo en fumadores, exfumadores y no fumadores. *Horiz Med 2016*, 6. Obtenido de https://n9.cl/jibne

Espinosa-Zúñiga, J. J. (2020). Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito. *Ingeniería, investigación y tecnología*, 16. Obtenido de https://n9.cl/mmleg

FRITZ, M. J. (s.f.). Métodos de Regularizaci´on para la Selección de Variables Aplicados a la Predicci´on del Riesgo de Padecer disfunci´on motora en Adultas Mayores Activas de la Ciudad de Valdivia. *Tesis Para Optar al Grado de Mag´ıster en Estad´ıstica.* Universidad de Concepci´on, Chile. Obtenido de https://n9.cl/70bhe

González, R., Barrientos, A., Toapanta, M., & Cerro, J. d. (2017). Aplicación de las Máquinas de Soporte Vectorial (SVM) al diagnóstico clínico de la Enfermedad de Párkinson y el Temblor Esencial . *Revista Iberoamericana de Automática e Informática Industrial*, 394-405. Obtenido de https://n9.cl/w91zd

GÓMEZ, R. (2016). GASTO EN SALUD DE PACIENTES DIAGNOSTICADOS CON EPOC, INCLUIDOS EN UN PROGRAMA DE ATENCIÓN DOMICILIARIA DE UNA IPS DE PEREIRA. TESIS DE POSGRADO. UNIVERSIDAD TECNOLÓGICA DE PEREIRA, PEREIRA. Obtenido de https://n9.cl/dfyyvj

Instituto Nacional de Estadística e Informática. (Mayo de 2023). Perú: Enfermedades transmisibles y no transmisibles 2022. Obtenido de INEI: https://n9.cl/890jt

JuliaReya, M.´.-R.-P., & LeonorVarela-Lemaa, h.´.-V. (2017). Mortalidad atribuida al consumo de tabaco en las comunidades autonomas de España. *Sociedad Española Cardiologa*, 9. Obtenido de https://n9.cl/jjfaz

Luis Pinillos A1, M. Q. (2005). TABAQUISMO:UN PROBLEMA DE SALUD PÚBLICA EN EL PERÚ. *Rev Peru Med Exp Salud Publica*, 7. Obtenido de https://n9.cl/dz9em

Martínez, R. E., Ramírez, N. C., Mesa, H. G., Suárez, I. R., Trejo, M. d., & Morales, P. P. (2009). Árboles de decisión como herramienta en el diagnóstico. *Instituto de Ciencias de la Salud, Universidad Veracruzana, Xalapa, Veracruz, México*, 6. Obtenido de https://n9.cl/ld4fy

médicos de la Sociedad Americana Contra El Cáncer. (23 de Junio de 2022). *American cancer Socity.* Obtenido de Por qué la gente comienza a fumar y por: https://n9.cl/rd9hmq

MINSA. (15 de junio de 2018). *FUMADORES PASIVOS INHALAN TRES VECES MÁS NICOTINA Y ALQUITRAN QUE LOS FUMADORES ACTIVOS*. Obtenido de MINSA: https://n9.cl/5zpe3b

Moreno, C. (31 de Junio de 2017). Cuando el humo de los cigarrillos también daña a la economía. *GESTIÓN*, pág. 1. Obtenido de https://n9.cl/3clp7

Núñez, A. (01 de Noviembre de 2019). El Comercio. *Cigarrillo electrónico en el Perú: ¿eficaz método contra la adicción o un riesgo?*, pág. 1. Obtenido de https://n9.cl/gk3fu

OCAMPO, E. M., GIRALDO, D. A., & ISAZA, H. S. (2006). PRONÓSTICO DE VENTAS USANDO REDES NEURONALES. *Scientia et Technica Año X, No 26*, 7. Obtenido de https://n9.cl/w0afs

Olabe, X. B. (2003). *REDES NEURONALES ARTIFICIALES Y SUS APLICACIONES.* Bilbao: Escuela Superior de Ingeniería de Bilbao, EHU. Obtenido de https://n9.cl/wlpbs

OPS. (Mayo de 2023). Control del tabaco. Obtenido de OPS: https://n9.cl/b5pag

Pichon-Riviere, A., Bardach, A., & Augustovski, F. (2016). Impacto económico del tabaquismo en los sistemas de salud de América Latina: un estudio en siete países y su extrapolación a nivel regional. *ÚMERO TEMÁTICO SOBRE ECONOMÍA DEL CONTROL DEL TABACO EN LAS AMÉRICAS*, 9. Obtenido de https://n9.cl/ygoge

Puga, J. L. (2009). MODELOS PREDICTIVOS EN ACTITUDES EMPRENDEDORAS:ANÁLISIS COMPARATIVO DE LAS CONDICIONES DE EJECUCIÓN DE LAS REDES BAYESIANAS Y LA REGRESIÓN LOGÍSTICA . *TESIS DOCTORAL.* Universidad de Almería, Almería. Obtenido de https://n9.cl/soyqj1

Radovic, F. (23 de Enero de 2019). Radio Exitosa | Flavia Radovic: “El Perú gasta 2,500 millones al año en enfermos de tabaquismo”. (N. Lucar, Entrevistador) Obtenido de https://colat.org/flavia-radovic-el-peru-gasta-2500-millones-al-ano-en-enfermos-de-tabaquismo/

Rev Med Hered. (2020). Estado actual de las investigaciones y riesgos del uso de los cigarrillos electrónicos. *Rev Med Hered*, 81-82.

Sociedad Peruana de Neumología. (2016). *Guía Peruana de EPOC -2016.* Obtenido de https://docs.bvsalud.org/biblioref/2019/07/1006744/guia_peruana_epoc.pdf

Stead LF, K. P. (2016). Combined pharmacotherapy and behavioural interventions forsmoking cessation (Review). CochraneLibrary, 98. Obtenido de https://n9.cl/36wka8

Velásquez, J. D., Olaya1, Y., & Franco, C. J. (2010). PREDICCIÓN DE SERIES TEMPORALES USANDO MÁQUINAS. *Ingeniare. Revista chilena de ingeniería, vol. 18 Nº 1*, 12. Obtenido de https://n9.cl/9rg4h

Villarreal, E., Julián, Y., & Emma, V. (2018). Costo de la atención médica en pacientes con enfermedad pulmonar obstructiva crónica. *IMSS*. Obtenido de https://n9.cl/975t6

World Health organization. (24 de Mayo de 2022). *Tobacco*. Obtenido de World Health organization: https://n9.cl/crx2z

Zhang, R., Li, B., & Jiao, B. (2019). Application of XGboost Algorithm in Bearing Fault Diagnosis. *OP Conference Series: Materials Science and Engineering*, 6. Obtenido de https://n9.cl/6duru

Zhao, D. Z., & Wang, X. (2020). Association of Smoking and Smoking Cessation With Overall and Cause-Specific Mortality. *The Lancet*. Obtenido de https://n9.cl/iihww/