Universidad Nacional de Asunción
Facultad Politécnica

Maestría en Ciencias de la Computación

Análisis Exploratorio de las Relaciones Lineales y no Lineales de las Señales de Fotopletismografía Aplicando Análisis de Componentes Principales para la Estimación de la Presión Arterial

Carolina Elizabeth Villegas Colmán

Trabajo de Maestría presentado en conformidad a los
requisitos para obtener el grado de Máster en Ciencias de la Computación

San Lorenzo - Paraguay
Noviembre - 2021

Hoja de aprobación de Tesis

Análisis Exploratorio de las Relaciones Lineales y no Lineales de las Señales de Fotopletismografía Aplicando Análisis de Componentes Principales para la Estimación de la Presión Arterial

Carolina Elizabeth Villegas Colmán

Tesis de Maestría aprobada el 30 de noviembre de 2021 por los siguientes miembros del Jurado de Defensa:

Prof. Dr. Pedro E. Gardel Sotomayor (UCA - Alto Paraná).
Prof. Dr. Daniel Romero (FPUNA).
Prof. D.Sc. José Luis Vázquez Noguera (FPUNA), co-orientador.
Prof. Dra. Cynthia Emilia Villalba Cardozo (FPUNA), orientadora.

Prof. Dr. Horacio A. Legal Ayala
Coordinador Académico
Postgrado en Ciencias de la Computación
Facultad Politécnica
Universidad Nacional de Asunción

Prof. Dra. Cynthia Emilia Villalba Cardozo
Orientadora

Universidad Nacional de Asunción
Facultad Politécnica

Maestría en Ciencias de la Computación

Análisis Exploratorio de las Relaciones Lineales y no Lineales de las Señales de Fotopletismografía Aplicando Análisis de Componentes Principales para la Estimación de la Presión Arterial

Carolina Elizabeth Villegas Colmán

San Lorenzo - Paraguay
Noviembre - 2021

Universidad Nacional de Asunción
Facultad Politécnica

Maestría en Ciencias de la Computación

Análisis Exploratorio de las Relaciones Lineales y no Lineales de las Señales de Fotopletismografía Aplicando Análisis de Componentes Principales para la Estimación de la Presión Arterial

Carolina Elizabeth Villegas Colmán

Orientadores:

Dra. Cynthtia Emilia Villalba Cardozo
D.Sc. José Luis Vázquez Noguera

Trabajo de Maestría presentado en conformidad a los
requisitos para obtener el grado de Máster en Ciencias de la Computación

San Lorenzo - Paraguay
Noviembre - 2021

A mi familia,
y a mis mejores amigos.

AGRADECIMIENTOS

A Dios y a la Virgencita por bendecirme y permitirme llegar a la meta.

A mis padres, por guiarme y enseñarme siempre bajo los principios y valores de la honestidad, integridad, respeto e independencia.

A mis familiares y amigos, por sus demostraciones de apoyo incondicional, y sus consejos de superación ante las dificultades.

A mis tutores, la Dra. Cynthtia Emilia Villalba Cardozo y el D.Sc. José Luis Vázquez Noguera por sus enseñanzas, consejos, recomendaciones y orientaciones durante el desarrollo de esta tesis.

A los profesores, el M.Sc. Santiago Gómez Guerrero y el Dr. Miguel García Torres por sus colaboraciones y recomendaciones durante el desarrollo de esta tesis.

A mis compañeros, por el apoyo y la ayuda que me han brindado desde el inicio de la maestría.

Y al Consejo Nacional de Ciencia y Tecnología (CONACyT), por haberme dado la oportunidad de acceder a este programa de maestría.

Análisis Exploratorio de las Relaciones Lineales y no Lineales de las Señales de Fotopletismografía Aplicando Análisis de Componentes Principales para la Estimación de la Presión Arterial

Autor: Carolina Elizabeth Villegas Colmán

Orientadores: Dra. Cynthtia Emilia Villalba Cardozo
D.Sc. José Luis Vázquez Noguera

RESUMEN

Las señales de fotopletismografía en modelos de aprendizaje automático para la estimación de la presión arterial pueden verse afectadas cuando el conjunto de atributos es de alta dimensión. Reducir la dimensión de los atributos, eliminando aquellos atributos redundantes podrían ayudar a mejorar el rendimiento de los modelos predictivos. Este trabajo propone explorar las relaciones multivariadas lineales y no lineales de las señales de fotopletismografía. Esto se consigue aplicando análisis de componentes principales mediante la descomposición de las matrices de correlaciones de Pearson y Spearman como técnica de reducción de dimensionalidad. Los resultados muestran que la matriz de correlación de Spearman obtiene levemente una mayor varianza acumulada en los primeros componentes principales. Las predicciones de la presión arterial con los nuevos componentes principales generados para los modelos predictivos satisfacen los estándares de la Asociación para el Avance de la Instrumentación Médica.

Palabras Clave: Análisis de Componentes Principales, Aprendizaje Automático, Correlaciones de Pearson y Spearman, Presión Arterial.

Exploratory Analysis of Linear and Nonlinear Relationships of Photoplethysmography Signals Applying Principal Component Analysis for Blood Pressure Estimation

Author: Carolina Elizabeth Villegas Colmán

Advisors: Dra. Cynthtia Emilia Villalba Cardozo
D.Sc. José Luis Vázquez Noguera

ABSTRACT

Photoplethysmography signals in machine learning models for blood pressure estimation can be affected when the attribute set is high dimensional. Reducing the dimension of the attributes by discarding redundant attributes could help to improve the performance of predictive models. This work proposes to explore the linear and nonlinear multivariate relationships of photoplethysmography signals. This is achieved by applying principal component analysis by decomposing the Pearson and Spearman correlation matrices as a dimensionality reduction technique. The results show that the Spearman correlation matrix obtains slightly higher cumulative variance in the first principal components. The blood pressure predictions with the new principal components generated for the predictive models satisfy the standards of the Association for the Advancement of Medical Instrumentation.

Keywords: Blood Pressure, Machine Learning, Pearson and Spearman Correlations, Principal Component Analysis.

ÍNDICE

Página

LISTA DE FIGURAS

Página

LISTA DE TABLAS

Página

LISTA DE SÍMBOLOS

X	matriz de datos
i, k, n, p	variables numéricas
x_{ij}	elemento ubicado en la i-ésima fila y la j-ésima columna de la matriz X
Σ	matriz de covarianzas de X
λ	valor propio de la matriz de covarianzas Σ
Y_i	i-ésima combinación lineal de las variables iniciales
v'	vector propio en la dirección de máxima varianza
$Var(X_i)$	varianza de la i-ésima componente de X
$Var(Y_i)$	varianza de la i-ésima componente de Y
$Cov(Y_i, Y_k)$	covarianza entre la i-ésima y la k-ésima componente de Y
$\sum$	sumatoria
σ	varianza de la variable inicial
Λ	matriz de covarianzas de Y
$tr(\Lambda)$	traza de la matriz Λ
$tr(\Sigma)$	traza de la matriz Σ
Z	matriz de las variables estandarizadas de X
μ	media aritmética de las componentes de X
$V^{1/2}$	matriz diagonal de la desviación estándar
$E(Z)$	esperanza de la matriz Z
ρ	coeficiente de correlación de Pearson
ρ_s	coeficiente de correlación de Spearman
$\hat{Y}_i$	i-ésima combinación lineal de las variables estandarizadas
$\hat{\lambda}$	valor propio estandarizado de la matriz de correlaciones
$\hat{e}$	vector propio estandarizado en la dirección de máxima varianza
r	coeficiente de correlación de Pearson de una muestra
r_s	coeficiente de correlación de Spearman de una muestra
x_i	valor de x para el i-ésimo individuo
y_i	valor de y para el i-ésimo individuo
$\hat{x}$	media muestral de x
$\hat{y}$	media muestral de y
d_i	diferencia entre los rangos de x e y

CAPÍTULO 1

Introducción

La presión arterial (PA) es la presión ejercida por la sangre en las arterias durante la circulación sistémica [1]. La lectura de la PA se presenta mediante dos cifras:

- Presión arterial sistólica (PAS): es la presión máxima en la aorta cuando el corazón se contrae y expele la sangre del ventrículo izquierdo [1].
- Presión arterial diastólica (PAD): es la presión en los vasos sanguíneos cuando el corazón se encuentra en reposo entre los latidos [1].

Los rangos normales de la PA para un adulto sano en reposo es de 90/60 mmHg hasta 120/80 mmHg [2].

Cuando la medición de la PA se realiza en dos días diferentes y la PAS es superior o igual a 140 mmHg y la PAD es superior o igual a 90 mmHg en ambos días, entonces el diagnóstico es hipertensión [3].

La hipertensión es el principal factor de riesgo para algunas enfermedades cardiovasculares [1].

Se estima que más de mil millones de personas tienen hipertensión a nivel mundial, y con mayor prevalencia en países de ingresos bajos y medianos [1].

La medición exacta de la PA es fundamental para detectar y tratar adecuadamente la hipertensión [1].

La medición de la PA se puede realizar de forma continua, no invasiva y sin brazalete mediante la técnica de fotopletismografía [1].

La fotopletismografía (PPG, por sus siglas en inglés) es una señal circulatoria, no invasiva y está relacionada con el volumen pulsátil de la sangre en el tejido [4].

Como se observa en la Figura 1.1, la componente de corriente alterna (CA) de la señal de PPG es morfológicamente similar a la forma de onda de la presión arterial sanguínea (ABP, por sus siglas en inglés) [4].

Por lo cual, resulta intuitivo relacionarlas y buscar información sobre la PA a partir de la señal de PPG [4].

El término CA hace referencia a las pulsaciones en el fotopletismograma generadas por los latidos del corazón [4].

En la Figura 1.1 se comprimen varios segundos de adquisición en una sola imagen, y además se observan las fases de adquisición de las señales de PPG y ABP.

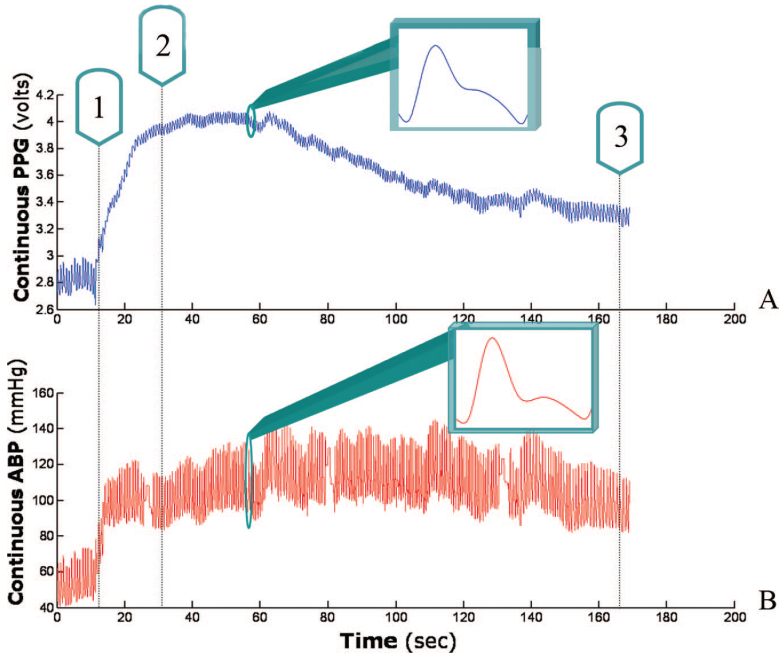

Figura 1.1: Comparación morfológica de las señales de PPG y ABP [4].

La propuesta es utilizar las señales de PPG con sus valores de referencias de la PA (ABP), implementando algoritmos de aprendizaje automático (ML, por sus siglas en inglés) para la estimación de la PA.

Para la estimación de la PA, las señales de PPG deben pasar básicamente por las siguientes etapas [5]:

1. Pre-procesamiento de las señales.
2. Extracción de características de las señales.
3. Reducción de la dimensionalidad de las características extraídas, y cálculo de los componentes principales (CPs).
4. Entrenamiento de los algoritmos de ML con los CPs y sus valores de referencias para la estimación de la PA.

A continuación se presentan algunos trabajos del estudio del arte que implementaron algoritmos de ML para la estimación de la PA en los últimos cinco años.

En [5], Mousavi et al. utilizaron los valores brutos de las señales de PPG con sus valores de referencias de la PA. Los mismos implementaron el análisis de componentes principales (PCA, por sus siglas en inglés) como técnica de reducción de dimensionalidad. Para la estimación de la PA se entrenaron los algoritmos de ML (decision tree regression (DTR) [6], support vector regression (SVR) [7], adaptative boosting regression [8] y random forest regression [9]).

En [10], Wang et al. desarrollaron un método para estimar la PA basado en las señales de PPG. El método multitaper (MTM, por sus siglas en inglés) se implementó para la extracción de características, y se usó una red neuronal artificial (ANN [11],por sus siglas en inglés) para la estimación de la PA.

En [12], Khalid et al. utilizaron las características más significativas de las señales de PPG (área del pulso, tiempo de aumento del pulso y ancho del 25 %) con sus valores de referencias de la PA. Se entrenaron los algoritmos de ML ((regression tree [6], multiple linear regression (MLR), y support vector machine (SVM) [7]) para la estimación de la PA.

Y en [13], Watanabe et al. mostraron la alta precisión y la gran ventaja de implementación del sensor de PPG como un potencial cambio de paradigma en la monitorización de la PA.

El artículo [5] sirvió de guía para el proceso de estimación de la PA implementando las señales de PPG con sus valores de referencias de la PA.

1.1. Justificación y Motivación

La etapa de extracción de características de las señales de PPG dan como resultado una matriz de datos de alta dimensión. Resulta una tarea difícil definir por simple inspección el comportamiento y los patrones de relaciones entre los atributos de la matriz.

Con frecuencia, los estudios de correlaciones preceden a análisis más complejos como el entrenamiento de algoritmos de ML (modelos predictivos) [14].

En esta tesis se propone la exploración de las relaciones multivariadas lineales y no lineales de las señales de PPG. Con esto se pretende analizar el comportamiento de las variables entre sí, y eliminar las variables redundantes de forma a reducir la matriz de alta dimensión.

Las exploraciones de las relaciones multivariadas se implementan dentro del PCA mediante la descomposición de las matrices de correlaciones de Pearson y Spearman como técnica de reducción de dimensionalidad. Esto permite encontrar nuevas variables que son combinaciones lineales del conjunto original de datos con una mínima pérdida de información. Estas nuevas variables se denominan componentes principales y se seleccionan de acuerdo a la varianza acumulada de los mismos.

Las contribuciones de esta tesis son: (1) reducir la dimensionalidad del conjunto de datos seleccionando los componentes principales de las señales de PPG con la aplicación del PCA mediante la descomposición de las matrices de correlaciones de Pearson y Spearman e (2) implementar los nuevos componentes principales para mejorar el rendimiento de los modelos predictivos en la estimación de la PA.

1.2. Objetivos

1.2.1. Objetivo General

- Reducir la dimensionalidad del conjunto de datos explorando las relaciones multivariadas lineales y no lineales de las señales de fotopletismografía a través del análisis de componentes principales.

1.2.2. Objetivos Específicos

- Explorar las relaciones multivariadas lineales y no lineales de las señales de fotopletismografía aplicando las correlaciones de Pearson y Spearman.
- Determinar las varianzas explicadas y acumuladas de los componentes principales aplicando análisis de componentes principales mediante la descomposición de las matrices de correlaciones de Pearson y Spearman.
- Seleccionar los componentes principales que retengan porcentajes de varianzas acumuladas con una mínima pérdida de información.
- Evaluar el rendimiento de los modelos predictivos entrenados con los componentes principales seleccionados para la estimación de la presión arterial.

1.3. Estructura del Documento

Este libro de tesis está estructurado como sigue:

En el Capítulo 2, se presenta el marco teórico sobre la técnica de análisis de componentes principales con sus formulaciones matemáticas.

En el Capítulo 3, se describen los procedimientos aplicados y los materiales utilizados para la estimación de la presión arterial.

En el Capítulo 4, se muestran los resultados experimentales obtenidos de cada uno de los procesos implementados para la estimación de la presión arterial.

Y en el Capítulo 5, se presentan las conclusiones de las principales contribuciones del trabajo.

CAPÍTULO 2

Análisis de Componentes Principales

En este capítulo estudiaremos las bases fundamentales de la técnica de análisis de componentes principales junto con sus formulaciones matemáticas, los cálculos de los componentes principales, las varianzas explicadas y acumuladas, y las exploraciones de las relaciones multivariadas lineales y no lineales.

2.1. Introducción

La técnica de reducción de datos es otro tipo de transformación de predictores [15].

Estas técnicas de transformaciones se implementan para reducir el impacto de los valores atípicos, y así producir mejoras en el rendimiento de los modelos predictivos [15].

Se puede definir a los valores atípicos como muestras que se alejan de forma inusual de la corriente principal de los datos [15].

Los métodos para resolver los valores atípicos, y reducir la dimensión de los datos son de gran importancia [15].

Cuando se sospecha que una o más variables del conjunto de datos son valores atípicos, se recomienda asegurarse de que estos valores son científicamente válidos [15].

En esta tesis serán omitidas las demostraciones matemáticas pero se indicarán las referencias correspondientes.

2.2. Componentes Principales

El análisis de componentes principales (PCA, por sus siglas en inglés) es una técnica que se encarga de reducir la dimensionalidad de un conjunto de datos conservando la mayor variabilidad posible [16]. Es decir, el objetivo del PCA es encontrar nuevas variables que son combinaciones lineales del conjunto original de datos, que maximicen la varianza y que no estén correlacionadas entre sí [16]. Estas nuevas variables se denominan componentes principales (CPs) [16], y se obtienen en orden decreciente de importancia.

Las bases teóricas del PCA se remontan a 1901 con Pearson [17], y 1933 con Hotelling [18].

En la Figura 2.1 se presenta de forma gráfica el proceso de implementar el PCA sobre una matriz de alta dimensión.

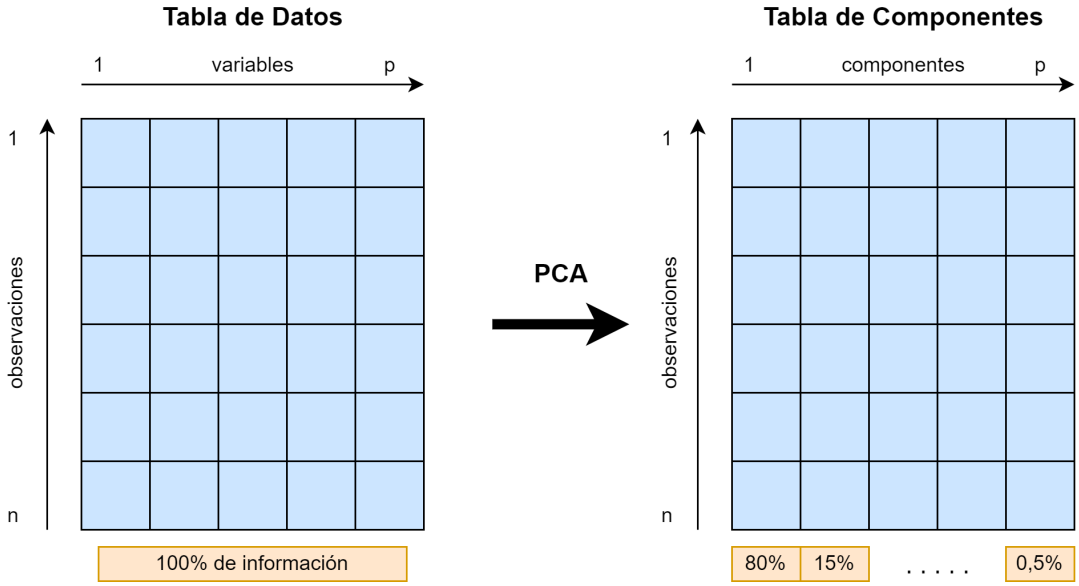

Figura 2.1: Análisis de Componentes Principales.

Como se observa en el lado izquierdo de la Figura 2.1, son p las variables necesarias para reproducir la variabilidad total del conjunto de datos, pero gran parte de esta variabilidad puede ser explicada por un número menor de k CPs (lado derecho) [19]. Siendo k un número menor a p.

Entonces, hay casi tanta información en los k componentes como en las p variables originales del conjunto de datos [19]. Y los k CPs pueden sustituir las p variables iniciales [19].

Por lo tanto, el conjunto original de datos que consistía en n observaciones sobre p variables, se reduce a un conjunto de datos de n observaciones sobre k CPs [19].

A continuación se presentan las formulaciones matemáticas para calcular los CPs, y las varianzas explicadas y acumuladas de los mismos.

2.2.1. Formulación Matemática

Sea un conjunto de datos con n observaciones sobre p variables numéricas. Estos valores definen p vectores $n-$dimensionales $X_1, X_2, \ldots, X_p$ o de forma equivalente, una matriz de datos X de dimensión $n \times p$ [16].

La matriz de datos X tiene la siguiente forma:

$$X = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \ldots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}$$

Desde el punto de vista algebraico, los CPs son combinaciones lineales de p variables aleatorias $X_1, X_2, \ldots, X_p$ [19].

Si seleccionamos los primeros dos vectores columnas, X_1 y X_2 de la matriz X, y graficamos cada uno de los componentes de estos vectores obtendremos una gráfica similar a la Figura 2.2. Los puntos de la gráfica se presentan de forma aleatoria.

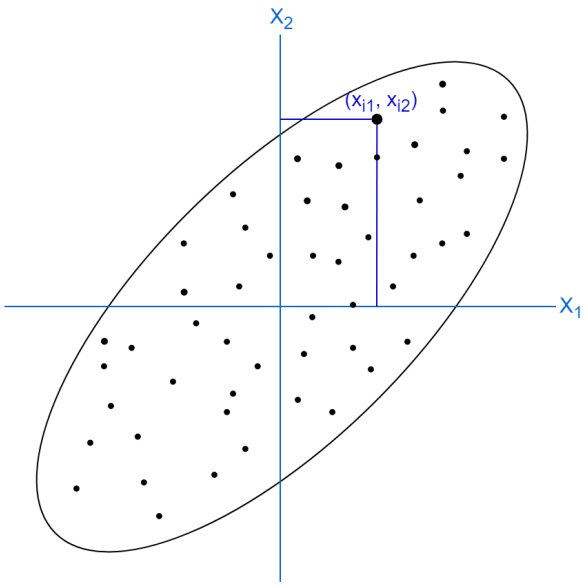

Figura 2.2: Proyección de los datos en dos dimensiones.

Como se observa en la Figura 2.2, con el PCA lo que se busca es la combinación lineal de las variables que recojan la mayor cantidad de información del conjunto de datos. Es decir, la dirección tal que al proyectar los puntos, la variabilidad sea la más grande posible.

Entonces, la dirección con máxima variabilidad del ejemplo corresponde a la dirección CP_1 proyectada en la Figura 2.3.

Luego, se busca la dirección que sea perpendicular a la primera y que recoja la mayor variabilidad restante, la cual corresponde a la dirección CP_2 de la Figura 2.3.

Como el análisis se realiza en dos dimensiones, entonces son dos los CPs que se determinan para este caso.

Para un caso general, se tendrán tantas CPs como variables tengamos inicialmente.

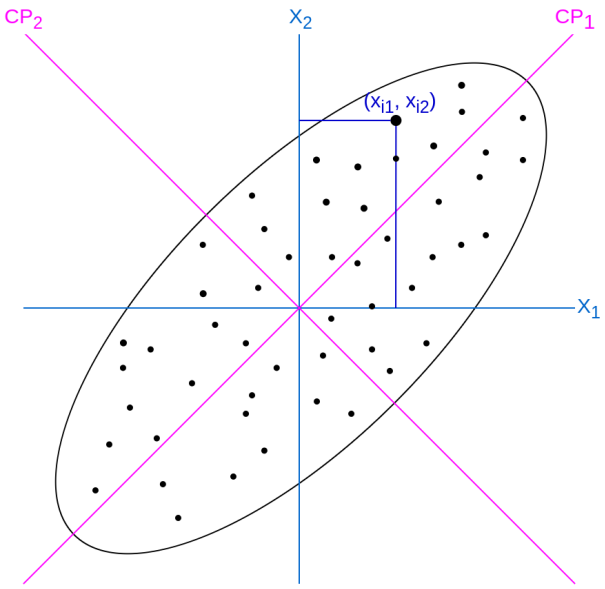

Figura 2.3: Identificación de las direcciones con máxima varianza.

Como se observa en la Figura 2.4, estas combinaciones lineales representan geométricamente la selección de un nuevo sistema de coordenadas obtenidos por la rotación del sistema original (Figura 2.3) [19].

Los nuevos ejes representan las direcciones con máxima variabilidad.

Las coordenadas de los puntos de la Figura 2.4 se calculan sobre los CPs, que son combinaciones lineales de las variables de partida.

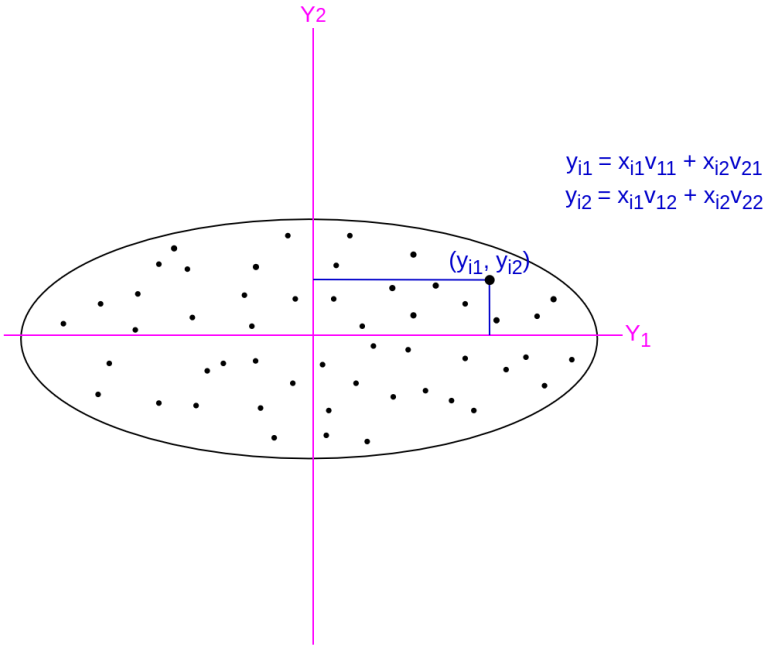

Figura 2.4: Rotación del sistema original de datos. Nuevos ejes de coordenadas con máxima variabilidad.

A continuación se presentan las formulaciones matemáticas de forma general.

Los CPs dependen únicamente de la matriz de covarianzas Σ con valores propios

$\lambda_1 \geq \lambda_2 \geq ...\lambda_p \geq 0$, o la matriz de correlaciones ρ (matriz estandarizada de Σ) de $X_1, X_2, ..., X_p$ [19].

Considere las siguientes combinaciones lineales:

$$\begin{aligned} Y_1 &= v_1'\mathbf{X} = v_{11}x_1 + v_{12}x_2 + \cdots v_{1p}x_p \\ Y_2 &= v_2'\mathbf{X} = v_{21}x_1 + v_{22}x_2 + \cdots v_{2p}x_p \\ \vdots &\quad \vdots \\ Y_p &= v_p'\mathbf{X} = v_{p1}x_1 + v_{p2}x_2 + \cdots v_{pp}x_p \end{aligned} \tag{2.1}$$

La varianza de la combinación lineal viene dada por:

$$Var(Y_i) = v_i'\Sigma v_i \quad i = 1, 2, ..., p \tag{2.2}$$

Y la covarianza entre las combinaciones lineales Y_i e Y_j viene dada por:

$$Cov(Y_i, Y_k) = v_i'\Sigma v_k \quad i, k = 1, 2, ..., p \tag{2.3}$$

Por lo cual, se busca la combinación lineal de las columnas de la matriz X con la máxima varianza.

Por lo tanto, los CPs se definen como [19]:

$$\begin{aligned} CP_1 &= \text{es la combinación lineal de } v_1'X \text{ que maximiza } Var(v_1'X) \text{ sujeto a} \\ &\quad \text{la restricción de que } v_1'v_1 = 1. \\ CP_2 &= \text{es la combinación lineal de } v_2'X \text{ que maximiza } Var(v_2'X) \text{ sujeto a} \\ &\quad \text{la restricción de que } v_2'v_2 = 1 \text{ y } Cov(v_1'X, v_2'X) = 0. \\ \vdots &\qquad\qquad \vdots \\ CP_i &= \text{es la combinación lineal de } v_i'X \text{ que maximiza } Var(v_i'X) \text{ sujeto a} \\ &\quad \text{la restricción de que } v_i'v_i = 1 \text{ y } Cov(v_i'X, v_k'X) = 0 \text{ para } k < i. \end{aligned} \tag{2.4}$$

La varianza total de los CPs se determina de la siguiente manera:

$$\sum_{i=1}^{p} Var(Xi) = tr(\Sigma) = tr(\Lambda) = \sum_{i=1}^{p} Var(Y_i)$$

Lo que implica que:

$$\begin{aligned} \text{Varianza total de la población} &= \sigma_{11} + \sigma_{22} + \ldots + \sigma_{pp} \\ &= \lambda_1 + \lambda_2 + \ldots + \lambda_p \end{aligned} \tag{2.5}$$

Es decir, la suma de las varianzas de las variables originales y la suma de las varianzas de las CPs son iguales.

Entonces, la proporción de la varianza total explicada por el CP_k es:

$$\frac{\lambda_k}{\lambda_1 + \lambda_2 + \ldots + \lambda_p} \quad k = 1, 2, ..., p \tag{2.6}$$

Los CPs se seleccionan de acuerdo a un porcentaje de varianza acumulada. Por ejemplo: Si los primeros 1, 2 o 3 CPs presentan entre el 80-90 % de la varianza total de la población, entonces estos componentes pueden reemplazar las p variables originales sin mucha pérdida de información [19].

2.3. Componentes Principales Mediante Variables Estandarizadas

A continuación se presenta el análisis de componentes principales (PCA) con variables estandarizadas. En donde la matriz de datos original se transforma realizando una descomposición espectral de las matrices de correlaciones de Pearson y Spearman por separado.

Las variables de entrada deben ser estandarizadas antes de aplicar el PCA.

Las variables de la matriz X se estandarizan si tienen diferentes unidades de medida, o si se desea que cada variable reciba el mismo peso en el análisis [20]. Para estandarizar las variables de X, se resta la media y se divide por la desviación estándar de cada componente como se observa a continuación:

$$\begin{aligned} Z_1 &= \frac{(X_1 - \mu_1)}{\sqrt{\sigma_{11}}} \\ Z_2 &= \frac{(X_2 - \mu_2)}{\sqrt{\sigma_{22}}} \\ \vdots & \quad \vdots \\ Z_p &= \frac{(X_p - \mu_p)}{\sqrt{\sigma_{pp}}} \end{aligned} \tag{2.7}$$

En notación matricial,

$$Z = (V^{1/2})^{-1}(X - \mu) \tag{2.8}$$

donde $V^{1/2}$ es la matriz diagonal de la desviación estándar.

Teniendo en cuenta que el $E(Z) = 0$ y $Cov(Z) = (V^{1/2})^{-1}\Sigma(V^{1/2})^{-1} = \rho$ [19].

Los CPs de Z pueden obtenerse a partir de los vectores propios de la matriz de correlación ρ de X [19].

Los análisis de la sección anterior se pueden aplicar con los coeficientes de correlaciones de Pearson (ρ) y Spearman (ρ_s) en lugar de Σ, pero con algunas simplificaciones ya que la varianza de cada Z_i es la unidad. Se utiliza la notación $\hat{Y}_i$ para referirse a la i-ésima CP, y $(\hat{\lambda}_i, \hat{e}_i)$ para referirse al par valor propio - vector propio de ρ [19].

Las combinaciones lineales de las variables estandarizadas vienen dadas por:

$$\begin{aligned} \hat{Y_1} &= \hat{e_{11}}Z_1 + \hat{e_{12}}Z_2 + \cdots \hat{e_{1p}}Z_p \\ \hat{Y_2} &= \hat{e_{21}}Z_1 + \hat{e_{22}}Z_2 + \cdots \hat{e_{2p}}Z_p \\ \vdots & \quad \vdots \\ \hat{Y_p} &= \hat{e_{p1}}Z_1 + \hat{e_{p2}}Z_2 + \cdots \hat{e_{pp}}Z_p \end{aligned} \tag{2.9}$$

El resto de los procedimientos y las interpretaciones siguen como se discutió en la sección anterior.

A continuación se presentan los dos coeficientes de correlaciones implementados para las exploraciones de las relaciones multivariadas lineales y no lineales del conjunto de datos.

2.3.1. Exploraciones Multivariadas Lineales y No Lineales

Antes de iniciar con las exploraciones de las relaciones multivariadas del conjunto de datos, debemos tener en cuenta la siguiente definición:

Correlación: es un método que se encarga de evaluar una posible asociación lineal bidireccional entre dos variables continuas [21]. Se mide mediante un estadístico denominado coeficiente de correlación, y representa la fuerza de la posible asociación lineal entre las variables analizadas [21]. Es adimensional y toma valores en el rango de -1 a +1 [21]. Los valores de las correlaciones pueden variar desde -1 (correlación negativa perfecta), pasando por 0 (sin correlación) a +1 (correlación positiva perfecta) [21].

Se pueden presentar diferentes tipos de relaciones entre las variables. Algunos ejemplos más representativos se presentan en la Figura 2.5.

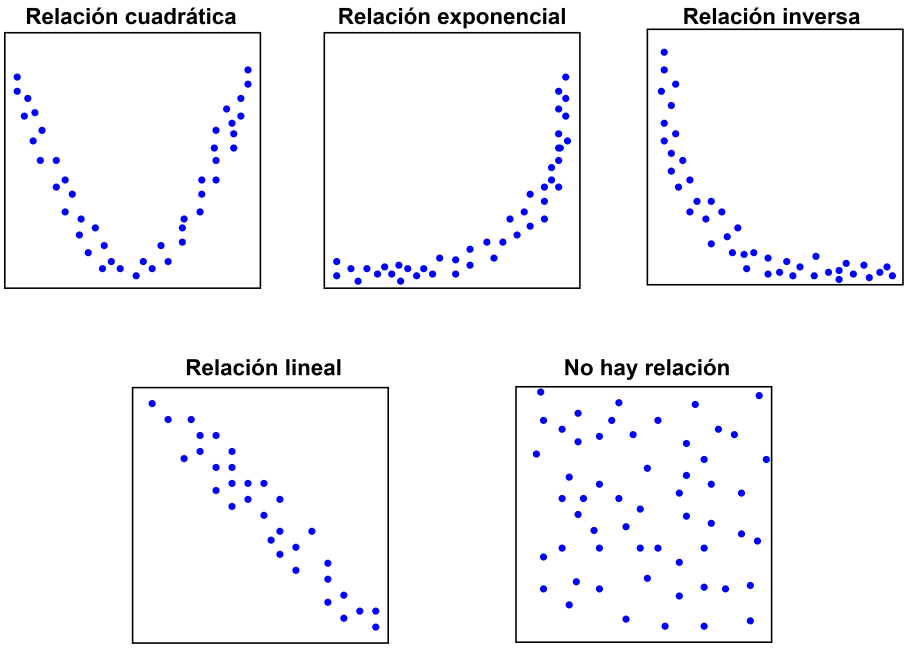

Figura 2.5: Ejemplos de las relaciones entre pares de variables.

Teniendo en cuenta que los datos de las señales de PPG son valores numéricos continuos, se aplicaron los siguientes coeficientes de correlaciones.

Coeficiente de Correlación de Pearson: se implementa cuando las dos variables estudiadas están distribuidas normalmente [21]. Se denota como ρ para un parámetro poblacional y como r para una muestra [21]. Este coeficiente de correlación se ve afectado por los valores extremos que pueden amortiguar o exagerar la fuerza de la relación [21].

La fórmula para calcular la correlación de Pearson entre las variables x e y de la muestra es:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{[\sum_{i=1}^{n}(x_i - \bar{x})^2][\sum_{i=1}^{n}(y_i - \bar{y})^2]}} \tag{2.10}$$

En donde las barras indican las medias muestrales de x e y, y x_i e y_i son los valores de x e y para el i-ésimo individuo.

La propiedad estadística de este coeficiente es una limitación intrínseca del PCA, que restringe su aplicabilidad a conjuntos de datos específicos con relaciones mayoritariamente lineales entre las entradas [22].

Coeficiente de Correlación de Spearman: es una medida no paramétrica de la dependencia estadística entre dos variables que cuantifica el grado de relación entre dos variables mediante una función monótona [22].

Este coeficiente se denota como ρ_s para un parámetro poblacional y como r_s para un estadístico muestral [21]. El valor absoluto de ρ_s describe la fuerza de la relación monótona [23]. Entonces, cuanto más se acerque este coeficiente al valor absoluto 0, más débil será la relación monótona entre las dos variables [23].

La fórmula para calcular la correlación de Spearman entre las variables x e y de la muestra es:

$$r_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \tag{2.11}$$

En donde n es el número de pares, y d_i es la diferencia entre los rangos de x e y.

La correlación de Spearman presenta menos sensibilidad a los valores atípicos, limitando a estos últimos a los valores de sus rangos [22].

Además la correlación de Spearman no es sensible a las desviaciones de la normalidad de las variables, y la linealidad entre pares de variables [22].

Por lo tanto, proporcionan un análisis completo de las correlaciones de primer orden a órdenes superiores (correlaciones basadas en las relaciones cuadradas, cúbicas o polinómicas de orden superior) [22].

En el siguiente capítulo se describen los procedimientos aplicados para la estimación de la PA.

CAPÍTULO 3

Metodología

En este capítulo se describen cada una de las etapas del proceso implementado para la estimación de la PA.

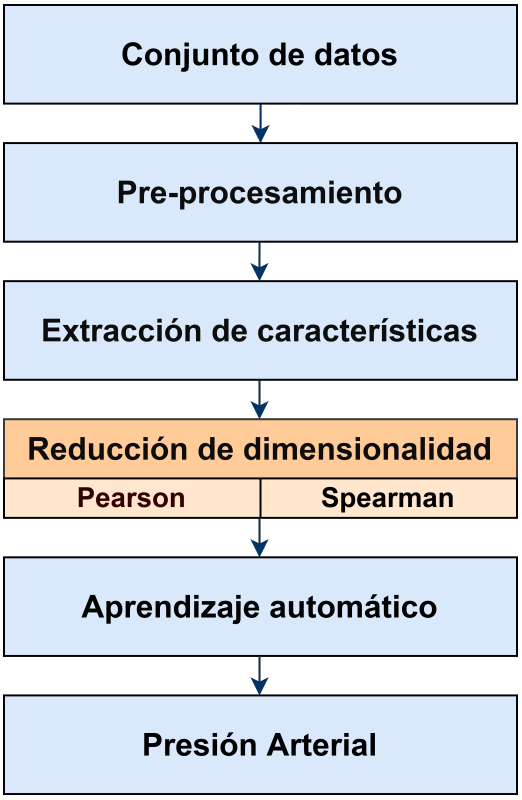

Figura 3.1: Diagrama de flujo propuesto para la estimación de la PA.

En la Figura 3.1 se observa la propuesta de esta tesis para la estimación de la PA.

A continuación se resumen los procesos implementados para la estimación de la PA:

1. Obtención y utilización de los segmentos de las señales de PPG y los valores de referencias de la PA adquiridos de una Base de Datos de libre acceso.
2. Pre-procesamiento de los segmentos de las señales de PPG, que incluyen filtro y normalización de las señales.
3. Extracción de características de los segmentos de las señales de PPG.
4. Reducción de la dimensionalidad de las características extraídas de las señales de PPG, y cálculo de los CPs.
5. Entrenamiento de los algoritmos de ML con los CPs y sus valores de referencias para la estimación de la PA.

3.1. Conjunto de Datos

Los registros multiparamétricos de las señales de PPG con sus valores de referencias de la PA (ABP) se obtuvieron de la Base de Datos MIMIC (Multiparameter Intelligent Monitoring in Intensive Care), disponible en la plataforma de PhysioNet [24].

MIMIC es una base de datos abierta que contiene la colección de registros multiparamétricos de pacientes internados en la Unidad de Cuidados Intensivos (UCI) [24].

En la Figura 3.2 se observan los segmentos de PPG y ABP adquiridos desde la Base de Datos MIMIC.

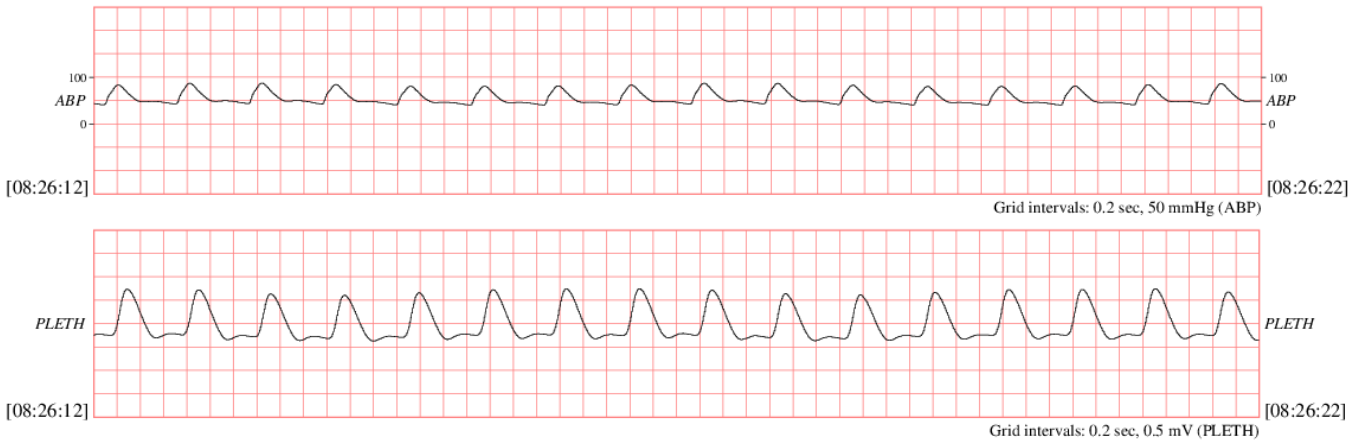

Figura 3.2: Señales de PPG y ABP de la Base de Datos MIMIC [24].

Se utilizaron 1145 segmentos que contenían las señales de PPG y los valores de referencias de la PA durante intervalos de 10 segundos [24].

La frecuencia de adquisición de las señales fue de 125 Hz con una precisión de 8 a 10 bits [24]. Las señales de PPG se registraron desde la yema de los dedos, y las señales del ABP desde la aorta de forma invasiva [24]. Estos datos informativos son proporcionados por la Base de Datos MIMIC.

3.2. Pre-procesamiento

El pre-procesamiento aplicado a las señales de PPG se compone de los siguientes pasos:

- Transformación matemática de las señales del dominio del tiempo al dominio de la frecuencia mediante la aplicación de la Transformada Rápida de Fourier (FFT, por sus siglas en inglés).
- Aplicación del filtro pasabanda de 0.4 a 8 Hz. Las frecuencias fuera del rango mencionado fueron reducidas a cero.
- Transformación matemática de las señales filtradas del dominio de la frecuencia al dominio del tiempo mediante la aplicación de la Transformada Inversa de Fourier.
- Normalización de las señales filtradas de PPG.

3.3. Extracción de Características

Los procedimientos implementados en esta etapa fueron los siguientes:

- Aplicación del algoritmo de detección automática de picos basada en la multiescala (AMPD, por sus siglas en inglés) para la determinación de los picos sistólicos en las señales de PPG como se puede observar en la Figura 3.3.
- Selección de los primeros dos picos sistólicos consecutivos de las señales de PPG.
- Extracción de los componentes de las señales de PPG entre los primeros dos picos sistólicos consecutivos determinados por el algoritmo AMPD como se observa en la Figura 3.4. Los componentes extraídos de cada uno de los segmentos de las señales de PPG fueron almacenados en una matriz de carga.
- Remuestreo de la frecuencia de las señales de PPG, disminuyendo el valor de la frecuencia para la estandarización de la cantidad total de componentes almacenados en la matriz de carga.
- Interpolación lineal para la estimación de los valores faltantes al estandarizar la matriz de carga.

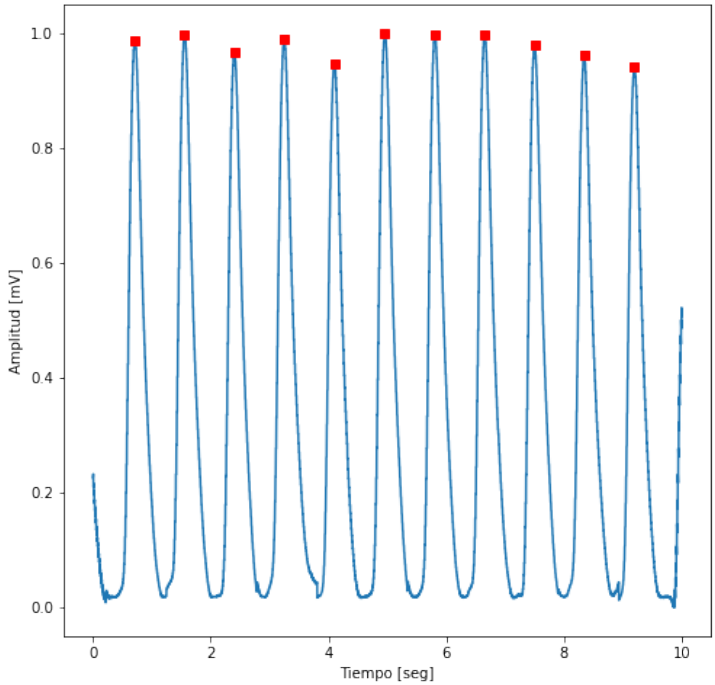

Figura 3.3: Detección de picos sistólicos en la señal de PPG mediante la aplicación del algoritmo AMPD.

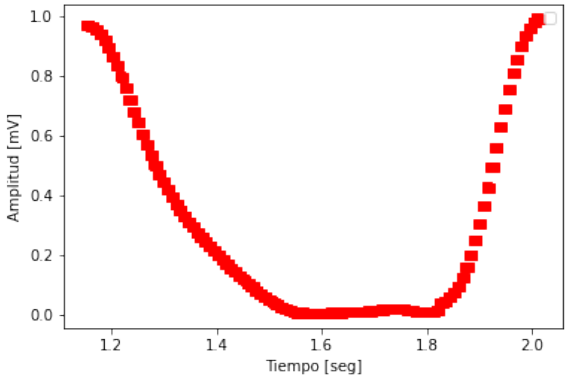

Figura 3.4: Extracción de características entre los primeros dos picos sistólicos de la señal de PPG.

3.4. Reducción de Dimensionalidad

En esta etapa se aplicó el análisis de componentes principales (PCA) como técnica de reducción de dimensionalidad, y los pasos seguidos fueron los siguientes:

- Estandarización de los datos de entrada.
- Cálculo de las matrices de correlaciones de Pearson y Spearman.
- Cálculo de los valores propios y los vectores propios a partir de las matrices de correlaciones de Pearson y Spearman.
- Cálculo de los componentes principales (CPs) teniendo en cuenta el procedimiento anterior.
- Selección de los CPs teniendo en cuenta la varianza acumulada de los mismos.

Luego de seleccionar la cantidad de CPs a tener en cuenta para la reducción de la dimensionalidad, fue necesario determinar los valores máximos y mínimos correspondientes a los valores de referencias de la PA. Esto se obtuvo a partir de las señales del ABP.

El valor máximo del ABP corresponde al valor de la presión arterial sistólica (PAS), y el valor mínimo corresponde al valor de la presión arterial diastólica (PAD).

A partir de estos últimos valores, se calculó el valor de la presión arterial media (PAM) de la siguiente manera:

$$PAM = \frac{PAS + 2PAD}{3} \tag{3.1}$$

Los valores de la PAS, la PAD y la PAM determinados anteriormente corresponden a los valores objetivos para la estimación de la PA.

Estos tres valores objetivos se ensamblaron por separado en matrices que contenían los CPs seleccionados anteriormente, y teniendo en cuenta las correlaciones aplicadas.

Se debe tener en cuenta que para las estimaciones de la PAS, la PAD y la PAM, el proceso se realiza por separado para cada uno de los valores objetivos.

En cada uno de los casos, el 70 % de los datos fueron utilizados para el entrenamiento de los modelos predictivos, y el 30 % restante para las pruebas de las estimaciones de la PAS, la PAD y la PAM.

3.5. Aprendizaje Automático

El aprendizaje automático o aprendizaje de máquinas, traducido en el idioma inglés como Machine Learning (ML), es el campo de estudio que otorga a las computadoras la capacidad de aprender sin ser explícitamente programadas [25].

En la Figura 3.5, se presenta la construcción de un modelo predictivo en el aprendizaje automático.

El computador observa los datos de entrada (ejemplo), aplica el algoritmo de aprendizaje automático y a partir de esto contruye un modelo de regresión. Entonces, crea una hipótesis acerca de los datos manejados, lo cual le permite resolver problemas (dar respuestas) con nuevos casos de entrada.

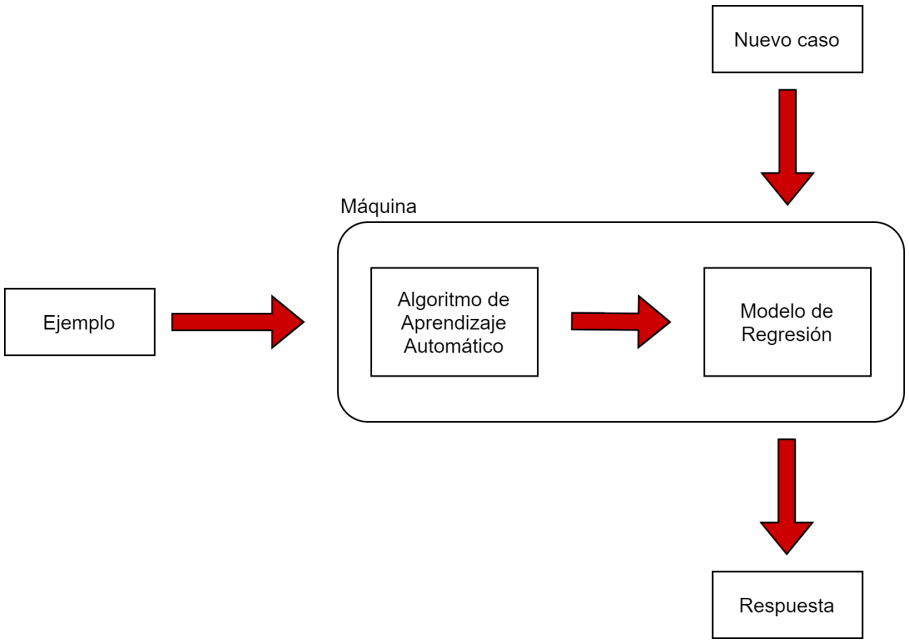

Figura 3.5: Construcción de un modelo predictivo en el aprendizaje automático.

3.5.1. Algoritmos de Aprendizaje Automático

A continuación se presenta una descripción básica de los algoritmos de ML implementados para la estimación de la PA.

Decision Tree Regression (DTR). El método implementado está basado en el modelo CART (classification and regression tree) introducido por Breiman et. al (1984) en [6]. Este método genera un árbol binario que divide el conjunto de datos a partir de un nodo raíz basado en la condicional sí/no de las variables independientes. Los subnodos creados son más puros que el nodo raíz. Durante este proceso, se buscan candidatos para alcanzar la división óptima que da lugar a un árbol con alta pureza. El proceso de partición se repite hasta alcanzar la condición de parada asignada previamente [26].

Random Forest Regression (RFR). Fue propuesto por Breiman (2001) en [9]. Es una combinación de árboles predictores. Cada árbol depende de los valores de un vector aleatorio muestreado independientemente y con la misma distribución para todos. Se aplica una selección aleatoria de características para dividir cada nodo. Una vez generado un gran número de árboles, se utiliza un esquema de votación para obtener el resultado final. En el caso de la regresión, la predicción final está representada por la media de las predicciones de cada árbol inducido.

Adaptative Boosting Regression (AdaBoost.R). Es un método de aprendizaje por conjuntos propuesto por Freund et al. (1997) en [8]. Combina múltiples aprendices de base. El modelo puede mejorar la precisión de los aprendices débiles cambiando la distribución de los pesos de las muestras. Se asignan pesos mayores a las muestras que clasificaron con resultados erróneos después de una iteración de entrenamiento. Estas reciben más atención durante el proceso de entrenamiento del siguiente aprendiz base. Los resultados finales del modelo se generan mediante combinación ponderada de todos los aprendices de base.

Support Vector Regression (SVR). Este método fue propuesto por Vapnik (1995) en [7]. Se formula como un problema de optimización, donde se construye una función multiobjetivo a partir de la función de pérdida y las propiedades geométricas del tubo. El hiperplano se representa en términos de vectores de soporte, que son muestras de entrenamiento que se encuentran fuera del límite del tubo. Los vectores de soporte son las instancias más influyentes que afectan a la forma del tubo. Este tubo reformula el problema de optimización para encontrar el que mejor se aproxime a la función del valor continuo, al tiempo que equilibra la complejidad del modelo y el error de predicción.

A continuación se presenta brevemente el lenguaje de programación implementado para el proceso de estimación de la PA con sus principales bibliotecas y funciones.

3.6. Lenguaje de Programación

En esta tesis se utilizó el lenguaje de programación Python para implementar la metodología propuesta. Este lenguaje fue propuesto por Guido van Rossum [27]. Se eligió Python porque es un lenguaje de programación potente y de fácil aprendizaje. Tiene estructuras de datos eficientes de alto nivel con un enfoque simple pero eficaz para la programación orientada a objetos [27]. La sintaxis de Python y la tipificación dinámica, lo convierten en un lenguaje ideal para el desarrollo de aplicaciones en diversas áreas de estudio [27]. El intérprete de Python y las extensas bibliotecas están disponibles de forma gratuita en formato de código fuente para las principales plataformas en el sitio web de Python [27].

A continuación se detallan las bibliotecas y las funciones implementadas en Python de las principales etapas para la estimación de la PA.

3.6.1. Algoritmo de Pre-procesamiento de Datos

En la etapa de pre-procesamiento se aplicó el filtro pasabanda en las señales de PPG. Primero se implementó la Transformada Rápida de Fourier (FFT) para pasar las señales del dominio del tiempo al dominio de la frecuencia. De la biblioteca «scipy» se importó la función «fftpack» para la implementación de la FFT. En el dominio de la frecuencia se procedió a filtrar las señales dentro del rango de frecuencias comprendidas entre 0.4 a 8 Hz. Para volver al dominio del tiempo, de la biblioteca «scipy» se importó la función «ifft» para la aplicación de la inversa de la FFT. Y finalmente se realizó la normalización de las señales de PPG.

3.6.2. Algoritmo de Extracción de Características

En la etapa de extracción de características se implementó el algoritmo de detección automática de picos basada en la multiescala (AMPD). De la biblioteca «pyampd.ampd» se importó la función «find_peaks», que se encargó de detectar los picos sistólicos dentro de las señales de PPG. Posteriormente se seleccionaron los primeros dos picos sistólicos, y se extrajeron todos los componentes de las señales entre estos primeros dos picos sistólicos.

3.6.3. Algortimo de Reducción de Dimensionalidad

En la etapa de reducción de dimensionalidad se implementó el análisis de componentes principales (PCA). De la biblioteca «sklearn» se implementó la clase «sklearn.decomposition.PCA». El hiperparámetro configurado fue la cantidad de componentes a seleccionar para la reducción de datos. Se configuraron para 2, 3, 5 y 8 componentes principales.

Las correlaciones de Pearson y Spearman que se aplicaron a la matriz de datos estandarizada se realizaron con la biblioteca «pandas», implementando las funciones «.corr(method=pearson)» y «.corr(method=spearman)».

3.6.4. Algoritmos de Aprendizaje Automático

Se implementaron cuatro algoritmos de aprendizaje automático que fueron importados de la biblioteca «sklearn».

Decision Tree Regression (DTR): para la implementación de este algoritmo se importó **DecisionTreeRegressor** de la biblioteca «sklearn.tree» y los hiperparámetros se configuraron por defecto.

Random Forest Regression (RFR): para la implementación de este algoritmo se importó **RandomForestRegressor** de la biblioteca «sklearn.ensemble» y los hiperparámetros se configuraron por defecto.

Adaptative Boosting Regression (AdaBoost.R): para la implementación de este algoritmo se importó **AdaBoostRegressor** de la biblioteca «sklearn.ensemble». Los hiperparámetros se configuraron por defecto excepto «n_estimators» que se configuró para 500 estimadores. Este hiperparámetro se configuró de acuerdo al artículo guía [5] que se seleccionó durante la revisión bibliográfica.

Support Vector Regression (SVR): para la implementación de este algoritmo se importó **SVR** de la biblioteca «sklearn.svm» y los hiperparámetros se configuraron por defecto.

3.6.5. Métricas de Evaluación de Regresión

Para la evaluación de los modelos de regresión se implementó la biblioteca «sklearn.metrics». Este módulo implementa varias funciones de pérdida, puntuación y utilidad para medir el rendimiento de la regresión.

Error absoluto medio (MAE, por sus siglas en inglés): la función «mean_absolute_error» calcula el error absoluto medio, una métrica de riesgo correspondiente al valor esperado de la pérdida por error absoluto.

Error cuadrático medio (MSE, por sus siglas en inglés): la función «mean_squared_error» calcula el error cuadrático medio, una métrica de riesgo correspondiente al valor esperado del error o pérdida al cuadrado (cuadrático).

Raíz del error cuadrático medio (RMSE, por sus siglas en inglés): en este caso se calcula la raíz cuadrada (con la función «sqrt» de la biblioteca «numpy») de la función «mean_squared_error» que calcula el error cuadrático medio.

Para obtener el error medio de la diferencia entre el valor objetivo y el valor estimado, y la desviación estándar del mismo, se implementó la biblioteca «numpy».

Error medio (ME, por sus siglas en inglés): la función «numpy.mean()» calcula el promedio de los elementos de la matriz dada. La diferencia entre los valores objetivos y estimados fueron almacenados en esta matriz analizada.

Desviación estándar (STD, por sus siglas en inglés:) la función «numpy.std()» calcula la desviación estándar de la matriz dada a lo largo del eje especificado.

A continuación se presentan los experimentos y los resultados obtenidos durante el proceso de estimación de la PA.

CAPÍTULO 4

Experimentos y Resultados

En este capítulo se describen los diferentes experimentos planteados, y los resultados obtenidos de los procedimientos aplicados para la estimación de la PA.

Para analizar el rendimiento de los modelos predictivos hemos considerado tres experimentos diferentes:

- En la primera parte (Capítulo 4.1), se realizaron las exploraciones de las relaciones multivariadas lineales y no lineales aplicando las correlaciones de Pearson (lineal) y Spearman (no lineal) al conjunto de datos. Esto se realizó con el objetivo de evaluar la fuerza y la dirección de las relaciones entre pares de variables. Ambos coeficientes se aplicaron para explorar las relaciones existentes y determinar la correlación estadística adecuada al conjunto de datos implementado.

- En la segunda parte (Capítulo 4.2), se determinaron los porcentajes de varianzas explicadas y acumuladas de los CPs tras la aplicación del PCA mediante la descomposición espectral de las matrices de correlaciones de Pearson y Spearman. Esto se determinó con el objetivo de reducir las dimensiones del conjunto de datos a unos cuantos CPs, teniendo en cuenta los porcentajes de las varianzas acumuladas de los mismos con una mínima pérdida de información.

- Y finalmente (Capítulo 4.3), se analizaron los resultados obtenidos de los modelos predictivos (DTR, RFR, AdaBoost.R y SVR) con los CPs seleccionados anteriormente y sus valores de referencias para las estimaciones de la PAS, la PAD y la PAM.

4.1. Exploraciones de las Relaciones Lineales y No Lineales

En esta etapa se aplicaron las correlaciones de Pearson (lineal) y Spearman (no lineal) al conjunto de datos resultante de la etapa de extracción de características. Estos coeficientes de correlaciones fueron implementados para evaluar la fuerza y la dirección de las relaciones lineales y no lineales entre pares de variables [15]. Para las variables que se distribuyen normalmente se recomienda aplicar la correlación de Pearson y en caso contrario, la correlación de Spearman [15]. Cabe resaltar que el coeficiente de correlación de Spearman es más robusto a los valores atípicos que el coeficiente de correlación de Pearson [15]. Las relaciones identificadas mediante los coeficientes de correlaciones deben interpretarse como asociaciones y no como relaciones causales [15].

4.1.1. Matrices de Correlaciones de Pearson y Spearman

En la Figura 4.1 se presentan las matrices de correlaciones de Pearson y Spearman obtenidas al aplicar sobre el conjunto de datos. Cada matriz tiene una dimensión de 618 x 618 componentes. Los resultados de las correlaciones se presentan en formato tipo mapas de calor para analizar de forma visual las relaciones existentes entre pares de variables teniendo en cuenta la gama de colores según sus indicadores. Es decir, entre más claro sea el área de observación significa que la correlación entre esas variables es alta, y esto va decrementando hasta llegar a un color oscuro, el cual nos indica que no hay correlaciones entre esas variables. En ambos casos, se pueden observar que gran parte de las matrices presentaron correlaciones de moderadas a altas. La determinación de las correlaciones en el conjunto de datos es esencial para aplicar posteriormente el PCA y reducir el conjunto de datos de alta dimensión.

4.2. Varianzas Explicadas y Acumuladas de los Componentes Principales

En la Tabla 4.1, se presentan los porcentajes de las varianzas explicadas y acumuladas de los primeros diez CPs. Estos fueron calculados aplicando PCA mediante la descomposición de las matrices de correlaciones de Pearson y Spearman. La varianza explicada de cada CP toma valores decrecientes, es decir, cada CP presenta menos varianza con respecto al anterior. Se observa que desde el primer CP fue levemente más alta la varianza acumulada aplicando la correlación de Spearman con respecto a la correlación de Pearson. Los CPs se ordenan en forma decreciente de importancia como se observa en la Tabla 4.1.

En base a los resultados de la Tabla 4.1, se optaron por aplicar los criterios de reducción teniendo en cuenta el 70 %, el 80 %, el 90 % y el 96 % de varianzas acumuladas de los CPs. Estos criterios se aplicaron con el fin de determinar la cantidad de componentes representativos de

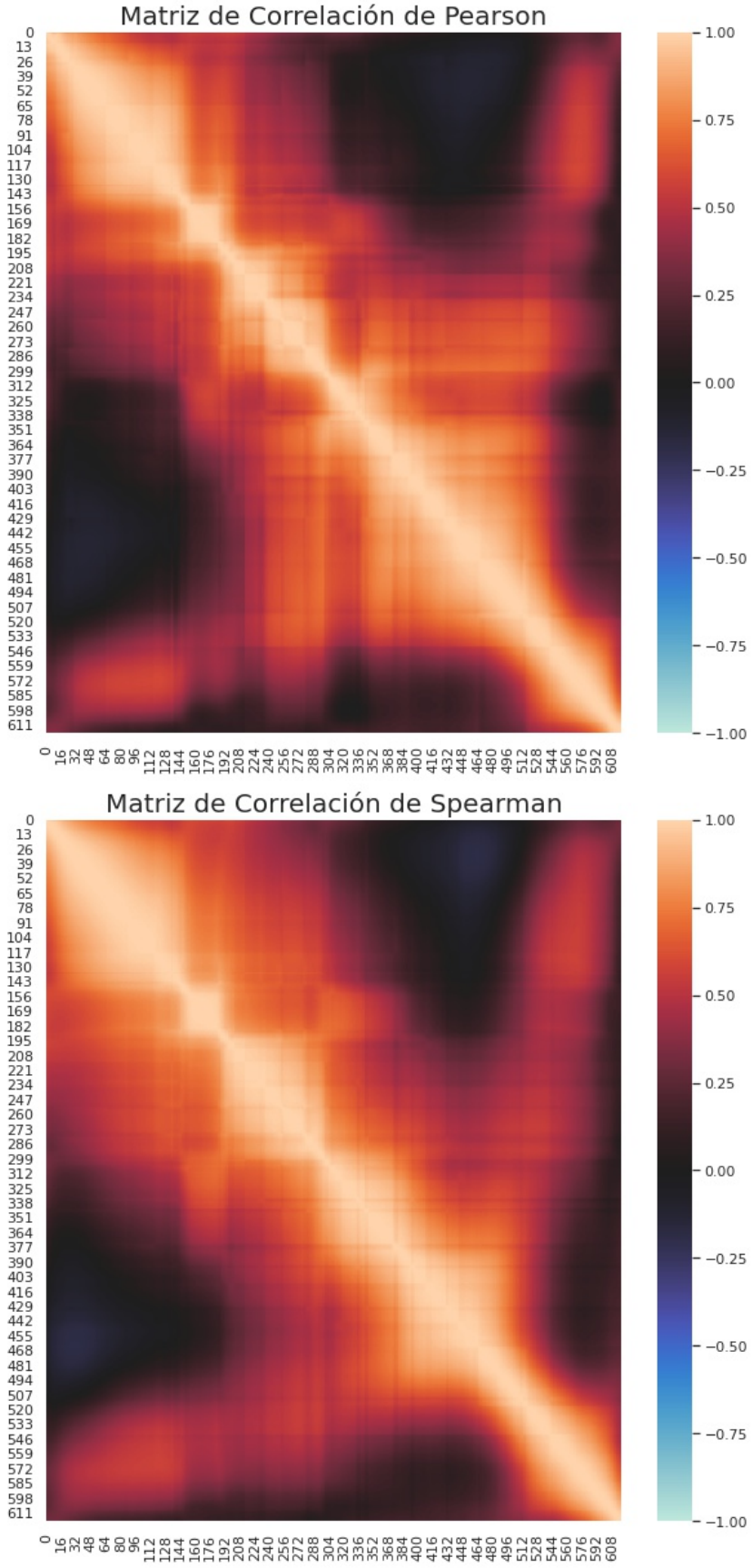

Figura 4.1: Matrices de Correlaciones de Pearson y Spearman.

Tabla 4.1: Varianzas explicadas y acumuladas de cada componente principal aplicando las matrices de correlaciones.

CP	PCA con Pearson		PCA con Spearman	
	% de varianza explicada	% varianza acumulada	% de varianza explicada	% varianza acumulada
1	47,93	47,93	49,45	49,45
2	25,13	**73,06**	23,95	**73,40**
3	10,15	**83,21**	10,62	**84,02**
4	3,89	87,10	4,26	88,28
5	3,43	**90,53**	3,43	**91,71**
6	2,54	93,07	2,44	94,15
7	1,87	94,94	1,72	95,87
8	1,31	**96,25**	1,17	**97,04**
9	0,81	97,06	0,68	97,72
10	0,70	97,76	0,42	98,14

la matriz de datos original, tal que mejore el rendimiento de los modelos predictivos para las estimaciones de la PAS, la PAD y la PAM.

Los 618 CPs obtenidos eran bastantes numerosos para implementarlos, por lo cual se eliminaron aquellos componentes que explicaban una proporción relativamente pequeña con respecto a la varianza total del conjunto. Esto se aplicó suponiendo que los primeros 2, 3, 5 y 8 CPs son componentes representativos del conjunto de datos original.

En efecto, se redujeron de 618 atributos originales a 2, 3, 5 y 8 CPs, los cuales retenían el 70 %, el 80 %, el 90 % y el 96 % de varianzas acumuladas con respecto al conjunto de datos original. Siendo estos nuevos componentes combinaciones lineales de los atributos originales no correlacionados entre sí y con una mínima pérdida de información.

4.3. Rendimiento de los Modelos Predictivos para la Estimación de la Presión Arterial

En las Tablas 4.2, 4.3 y 4.4, se presentan los resultados obtenidos de las métricas de evaluación de los modelos predictivos. Se implementaron varias funciones de pérdida, puntuación y utilidad para la evaluación de los modelos predictivos.

Tener en cuenta que y_i es el valor objetivo y x_i es el valor estimado de la iésima muestra, y n es el número de muestras. Las métricas de regresión aplicadas fueron:

- **Error medio (ME, por sus siglas en inglés)**: es la diferencia promedio entre el valor objetivo y el valor estimado por el modelo [28].

$$ME = (\frac{1}{n}) \sum_{i=1}^{n} (y_i - x_i) \tag{4.1}$$

- **Desviación estándar (STD, por sus siglas en inglés)**: es la desviación estándar del vector que contiene la diferencia entre el valor objetivo y el valor estimado a lo largo del eje especificado [28].

$$STD = \sqrt{(\frac{1}{n})\sum_{i=1}^{n}((y_i - x_i) - mean)^2} \tag{4.2}$$

- **Error absoluto medio (MAE, por sus siglas en inglés)**: es el promedio de la diferencia absoluta entre el valor objetivo y el valor estimado por el modelo. MAE es una puntuación lineal que significa que todas las diferencias individuales se ponderan por igual [15].

$$MAE = (\frac{1}{n})\sum_{i=1}^{n}|y_i - x_i| \tag{4.3}$$

- **Error cuadrático medio (MSE, por sus siglas en inglés)**: es el promedio de la diferencia al cuadrado entre el valor objetivo y el valor estimado por el modelo. A medida que cuadra las diferencias, penaliza incluso un pequeño error que conduce a una sobreestimación de cuán malo es el modelo [15].

$$MSE = (\frac{1}{n})\sum_{i=1}^{n}(y_i - x_i)^2 \tag{4.4}$$

- **Raíz del error cuadrático medio (RMSE, por sus siglas en inglés)**: es la raíz cuadrada del promedio de la diferencia al cuadrado entre el valor objetivo y el valor estimado por el modelo. Los errores se cuadran primero antes del promedio, lo que representa una penalización alta en errores grandes. Esto implica que RMSE es útil cuando no se desean errores grandes [15].

$$RMSE = \sqrt{(\frac{1}{n})\sum_{i=1}^{n}(y_i - x_i)^2} \tag{4.5}$$

Para validar la exactitud clínica de los dispositivos de medición de la PA no invasiva se disponen de protocolos estandarizados.

Para la Asociación para el Avance de la Instrumentación Médica (AAMI, por sus siglas en inglés) se tienen las siguientes acotaciones: la diferencia media estimada frente a las mediciones de referencias de la PA deben ser ME $\leq$ 5 mmHg con una STD $\leq$ 8 mmHg para la PAS y la PAD [28].

El número de participantes debe ser de 85 o más [28], y la base de datos MIMIC incluye datos registrados de más de 90 pacientes de UCI [24].

Tabla 4.2: Métricas de evaluación de los modelos predictivos para la estimación de la PAS.

PAS						
ML	TRD	Nº	ME±STD	MAE	MSE	RMSE
DTR	PCA(PEARSON)	2	-0,10±11,15	1,70	124,43	11,15
	PCA(SPEARMAN)	2	0,19±8,52	0,94	72,55	8,52
	PCA(PEARSON)	3	0,23±9,94	1,30	98,85	9,94
	PCA(SPEARMAN)	3	1,26±14,43	2,54	209,83	14,49
	PCA(PEARSON)	5	0,24±7,53	0,77	56,83	7,54
	PCA(SPEARMAN)	**5**	**0±0,09**	**0,01**	**0,01**	**0,09**
	PCA(PEARSON)	8	0,47±6,14	0,48	37,94	6,16
	PCA(SPEARMAN)	8	-0,30±4,91	0,35	24,21	4,92
	NO APLICA	all	-0,53±6,62	0,65	44,13	6,64
RFR	PCA(PEARSON)	2	-0,12±7,96	1,49	63,41	7,96
	PCA(SPEARMAN)	2	0,25±6,25	0,99	39,14	6,26
	PCA(PEARSON)	3	-0,07±5,41	1,07	29,31	5,41
	PCA(SPEARMAN)	3	0,41±6,49	1,38	42,24	6,50
	PCA(PEARSON)	5	-0,01±6,08	1,16	36,91	6,08
	PCA(SPEARMAN)	5	0,30±4,71	1,16	22,28	4,72
	PCA(PEARSON)	8	0 ±4,34	0,93	18,86	4,34
	PCA(SPEARMAN)	8	0,49±6,01	1,39	36,33	6,03
	NO APLICA	all	-0,02±3,70	0,80	13,69	3,70
ADA	PCA(PEARSON)	2	-0,80±8,31	2,23	69,62	8,34
	PCA(SPEARMAN)	2	-0,69±7,31	2,22	53,94	7,34
	PCA(PEARSON)	3	-0,26±4,29	1,16	18,44	4,29
	PCA(SPEARMAN)	3	0,45±3,50	1,23	12,46	3,53
	PCA(PEARSON)	5	-0,30±5,59	1,23	31,36	5,60
	PCA(SPEARMAN)	5	-0,04±2,37	0,84	5,62	2,37
	PCA(PEARSON)	8	-0,25±3,26	1,10	10,66	3,27
	PCA(SPEARMAN)	8	0,14±1,73	0,78	3,01	1,73
	NO APLICA	all	-0,70±2,08	1,20	4,81	2,19
SVR	PCA(PEARSON)	2	0,19±9,13	1,40	83,39	9,13
	PCA(SPEARMAN)	2	0,19±9,13	1,40	83,40	9,13
	PCA(PEARSON)	3	0,20±9,13	1,39	83,42	9,13
	PCA(SPEARMAN)	3	0,20±9,13	1,39	83,42	9,13
	PCA(PEARSON)	5	0,20±9,13	1,38	83,41	9,13
	PCA(SPEARMAN)	5	0,20±9,13	1,38	83,41	9,13
	PCA(PEARSON)	8	0,20±9,13	1,38	83,39	9,13
	PCA(SPEARMAN)	8	0,20±9,13	1,38	83,39	9,13
	NO APLICA	all	0,20±9,13	1,38	83,40	9,13

Tabla 4.3: Métricas de evaluación de los modelos predictivos para la estimación de la PAD.

PAD						
ML	TRD	Nº	ME±STD	MAE	MSE	RMSE
DTR	PCA(PEARSON)	2	-0,54±12,74	2,42	162,71	12,76
	PCA(SPEARMAN)	2	-0,60±5,64	0,62	32,13	5,67
	PCA(PEARSON)	3	0±7,90	0,95	62,48	7,90
	PCA(SPEARMAN)	3	0,42±9,96	1,36	99,38	9,97
	PCA(PEARSON)	5	0,34±8,95	1,30	80,23	8,96
	PCA(SPEARMAN)	5	-0,49±7,58	1,10	57,76	7,60
	PCA(PEARSON)	8	0,89±12,77	1,81	163,97	12,80
	PCA(SPEARMAN)	8	-0,44±12	1,99	144,10	12
	NO APLICA	all	-0,64±11,33	1,88	128,74	11,35
RFR	PCA(PEARSON)	2	-0,24±8,98	2,79	80,75	8,99
	PCA(SPEARMAN)	2	0,05±7,86	2,30	61,72	7,86
	PCA(PEARSON)	3	0,37±8,31	2,92	69,20	8,32
	PCA(SPEARMAN)	3	0,47±8,84	2,84	78,34	8,85
	PCA(PEARSON)	5	0,44±7,02	2,62	49,50	7,04
	PCA(SPEARMAN)	**5**	**0,01±6,38**	**2,38**	**40,68**	**6,38**
	PCA(PEARSON)	8	0,40±6,40	2,39	41,13	6,41
	PCA(SPEARMAN)	8	-0,20±7,03	2,43	49,50	7,04
	NO APLICA	all	0,01±7,05	2,55	49,75	7,05
ADA	PCA(PEARSON)	2	7,14±15,45	11,99	289,62	17,02
	PCA(SPEARMAN)	2	4,86±15,24	10,14	255,81	15,99
	PCA(PEARSON)	3	8,76±15,80	13,42	326,16	18,06
	PCA(SPEARMAN)	3	2,55±15,16	7,81	236,36	15,37
	PCA(PEARSON)	5	8,86±15,82	13,09	328,87	18,13
	PCA(SPEARMAN)	5	8,72±14,49	12,62	286,03	16,91
	PCA(PEARSON)	8	10,96±13,64	15,14	306,11	17,50
	PCA(SPEARMAN)	8	7,65±13,38	11,43	237,54	15,41
	NO APLICA	all	10,68±11,49	13,83	246,18	15,69
SVR	PCA(PEARSON)	2	-2,72±17,41	4,28	310,57	17,62
	PCA(SPEARMAN)	2	-2,72±17,41	4,28	310,55	17,62
	PCA(PEARSON)	3	-2,73±17,41	4,28	310,55	17,62
	PCA(SPEARMAN)	3	-2,73±17,41	4,28	310,56	17,62
	PCA(PEARSON)	5	-2,73±17,40	4,27	310,28	17,61
	PCA(SPEARMAN)	5	-2,73±17,40	4,27	310,31	17,62
	PCA(PEARSON)	8	-2,73±17,40	4,26	310,24	17,61
	PCA(SPEARMAN)	8	-2,73±17,40	4,26	310,25	17,61
	NO APLICA	all	-2,72±17,40	4,27	310,25	17,61

Tabla 4.4: Métricas de evaluación de los modelos predictivos para la estimación de la PAM.

PAM						
ML	TRD	Nº	ME±STD	MAE	MSE	RMSE
DTR	PCA(PEARSON)	2	0,21±11,62	2,16	135,08	11,62
	PCA(SPEARMAN)	2	0,14±10,36	1,57	107,37	10,36
	PCA(PEARSON)	3	0,12±7,12	0,82	50,76	7,12
	PCA(SPEARMAN)	3	-0,10±8,68	1,56	75,29	8,68
	PCA(PEARSON)	5	0,22±10,49	1,52	110	10,49
	PCA(SPEARMAN)	**5**	**-0,23±2,99**	**0,24**	**9,02**	**3**
	PCA(PEARSON)	8	-0,41±4,91	0,66	24,29	4,93
	PCA(SPEARMAN)	8	-0,10±9,75	1,46	95,01	9,75
	NO APLICA	all	-0,54±8,74	1,40	76,68	8,76
RFR	PCA(PEARSON)	2	-0,25±7,43	2,18	55,24	7,43
	PCA(SPEARMAN)	2	0,08±6,87	1,83	47,21	6,87
	PCA(PEARSON)	3	0,26±6,45	2,18	41,64	6,45
	PCA(SPEARMAN)	3	0,21±6,43	2,15	41,36	6,43
	PCA(PEARSON)	5	0,32±6,04	1,98	36,59	6,05
	PCA(SPEARMAN)	5	-0,02±5,08	1,90	25,77	5,08
	PCA(PEARSON)	8	0,37±5,64	1,92	31,93	5,65
	PCA(SPEARMAN)	8	-0,06±5,25	1,80	27,56	5,25
	NO APLICA	all	-0,01±5,43	1,93	29,50	5,43
ADA	PCA(PEARSON)	2	0,58±12,88	8,18	166,31	12,90
	PCA(SPEARMAN)	2	3,58±12,46	9,20	168,12	12,97
	PCA(PEARSON)	3	4,40±11,63	9,31	154,66	12,44
	PCA(SPEARMAN)	3	3,13±12,47	7,03	165,20	12,85
	PCA(PEARSON)	5	9,90±11,74	12,43	235,83	15,36
	PCA(SPEARMAN)	5	5,44±10,16	8,25	132,87	11,53
	PCA(PEARSON)	8	10,28±10,48	12,72	215,42	14,68
	PCA(SPEARMAN)	8	5,90±12,28	8,89	185,63	13,62
	NO APLICA	all	9,96±9,05	12,28	181,09	13,46
SVR	PCA(PEARSON)	2	-1,75±13,58	3,25	187,54	13,69
	PCA(SPEARMAN)	2	-1,75±13,58	3,25	187,55	13,69
	PCA(PEARSON)	3	-1,75±13,58	3,25	187,61	13,70
	PCA(SPEARMAN)	3	-1,75±13,58	3,25	187,60	13,70
	PCA(PEARSON)	5	-1,75±13,58	3,24	187,40	13,70
	PCA(SPEARMAN)	5	-1,75±13,58	3,24	187,43	13,69
	PCA(PEARSON)	8	-1,74±13,58	3,23	187,36	13,69
	PCA(SPEARMAN)	8	-1,75±13,58	3,23	187,41	13,69
	NO APLICA	all	-1,75±13,58	3,24	187,43	13,69

Se realizan las siguientes acotaciones con respecto a las columnas de las Tablas 4.2, 4.3 y 4.4.

La columna ML corresponde a los algoritmos de aprendizaje automático aplicados (DTR, RFR, AdaBoost.R y SVR).

La columna TRD corresponde a la técnica de reducción de dimensionalidad aplicada (PCA-Pearson, PCA-Spearman o ninguna).

La columna Nº corresponde a la cantidad de CPs seleccionados para el entrenamiento de los modelos predictivos (2, 3, 5, 8 o el total de los atributos sin la aplicación de la TRD como paso previo). Las demás columnas corresponden a las métricas de evaluación aplicadas y analizadas anteriormente.

Para determinar los modelos predictivos con mejores rendimientos para las estimaciones de la PAS, la PAD y la PAM se analizaron las siguientes métricas de evaluación: ME±STD, MAE, MSE y RMSE.

Primero se seleccionaron los modelos predictivos con los menores valores de MSE y RMSE. A partir de estos, se seleccionaron los modelos predictivos con menor valor de MAE, y finalmente se verificaron si sus valores de ME±STD se encontraban en el rango del estándar de la AAMI para cada una de las estimaciones.

Los modelos predictivos con mejores rendimientos para las estimaciones de la PAS, la PAD y la PAM se observan en la Tabla 4.5. Y corresponden a los modelos predictivos con la aplicación previa de la técnica del PCA con la descomposición de la matriz de correlación de Spearman.

Tabla 4.5: Modelos predictivos con mejores rendimientos para las estimaciones de la PAS, la PAD y la PAM con respecto al estándar AAMI.

PA			
Estimación	Modelo predictivo y TRD	N°	ME±STD
PAS	DTR con PCA (Spearman)	5	0±0,09
PAD	RFR con PCA (Spearman)	5	0,01±6,38
PAM	DTR con PCA (Spearman)	5	-0,23±2,99
	AAMI (STANDARD)		($\leq$ 5mmHg) $\pm$ ($\leq$ 8mmHg)

Para analizar de forma gráfica los resultados obtenidos en la Tabla 4.5, utilizamos el gráfico de Bland-Altman.

El gráfico de Bland-Altman fue introducido por J. Bland y D. Altman, y se implementa para describir el acuerdo entre dos mediciones cuantitativas [29].

Con este gráfico, ellos establecieron un método para cuantificar la concordancia entre dos mediciones cuantitativas mediante la construcción de límites de concordancia [29].

Estos límites de concordancia se calculan utilizando la media y la desviación estándar de las diferencias entre dos mediciones [29].

El gráfico resultante es un diagrama de dispersión XY. El eje X representa la media de estas medidas ((A+B)/2), y el eje Y representa la diferencia entre las dos medidas emparejadas (A-B) [29].

Siendo A y B, las mediciones cuantitativas. En este caso, la medición A es el valor objetivo, y la medición B el valor estimado de la PA.

Se recomienda calcular el intervalo de confianza (IC) para observar la precisión de las estimaciones [29]. La precisión de la estimación depende de la cantidad de datos observados [29].

En particular, el IC del 95 % de la diferencia media muestra la magnitud de la diferencia sistemática [29].

J. Bland y D. Altman recomiendan que el 95 % de los puntos correspondientes a los datos, se sitúen dentro de ± 1.96SD de la diferencia media [29].

Los gráficos de Bland-Altman de las Figuras 4.2, 4.3 y 4.4, corresponden a los modelos predictivos seleccionados con mejores rendimientos de la Tabla 4.5, y los resultados obtenidos fueron los siguientes:

- Para la estimación de la PAS, el gráfico de la Figura 4.2 nos indica que los valores de la media de las mediciones se sitúan entre 80 a 125 mmHg aproximadamente. Se observa que para valores de 100 mmHg, la diferencia entre las mediciones presentaron variaciones que se encuentran dentro del intervalo de desviación con un 95 % de acuerdo, y con mínimos puntos fuera de este intervalo.

  Para valores por debajo de 100 mmHg, se observan que los mismos se encuentran sobre el margen cero con respecto a la diferencia media. Es decir, los valores objetivos y estimados son equivalentes en ese rango de medias (80-100 mmHg). También se pueden observar que sobre el margen cero se presentan 3 puntos bastantes aislados, para valores de media de 10, 125 y 180 mmHg.

  Los límites de concordancia para este caso fluctúan entre ±0,18 mmHg. El intervalo de confianza es bastante pequeño y con respecto a las muestras observadas, se obtuvieron buenos resultados.

- Para la estimación de la PAD, la gráfica de la Figura 4.3 nos indica que los valores de la media de las mediciones se sitúan entre 90 a 105 mmHg aproximadamente. Se observa que para valores dentro del rango mencionado anteriormente, se presentaron variaciones que se encuentran dentro del intervalo de desviación con un 95 % de acuerdo, y con algunos puntos fuera de este intervalo.

  Para valores por debajo de 75 mmHg, se observan que los mismos se encuentran en su mayoría fuera del intervalo de desviación esperado. Es decir, no hay concordancia entre los valores objetivos y estimados por debajo de 75 mmHg.

  Los límites de concordancia para este caso fluctúan entre ±12,5 mmHg. El intervalo de confianza es bastante amplio y con respecto a las muestras observadas, en su mayoría se obtuvieron buenos resultados pero también se presentaron errores significativos para valores por debajo de 75 mmHg.

- Para la estimación de la PAM, la gráfica de la Figura 4.4 nos indica que prácticamente la concordancia entre los valores objetivos y los valores estimados fue perfecta debido a que se sitúan sobre el margen cero. Pero esta concordancia se presenta desde valores muy bajos hasta 180 mmHg aproximadamente, y con un punto fuera de este intervalo de desviación. Es decir, no hay un rango limitado para los valores de la media de las mediciones.

  Los límites de concordancia para este caso fluctúan entre ±6 mmHg aproximadamente. El intervalo de confianza es amplio y con respecto a las muestras observadas, se obtuvieron buenos resultados pero las muestras analizadas son muy variables.

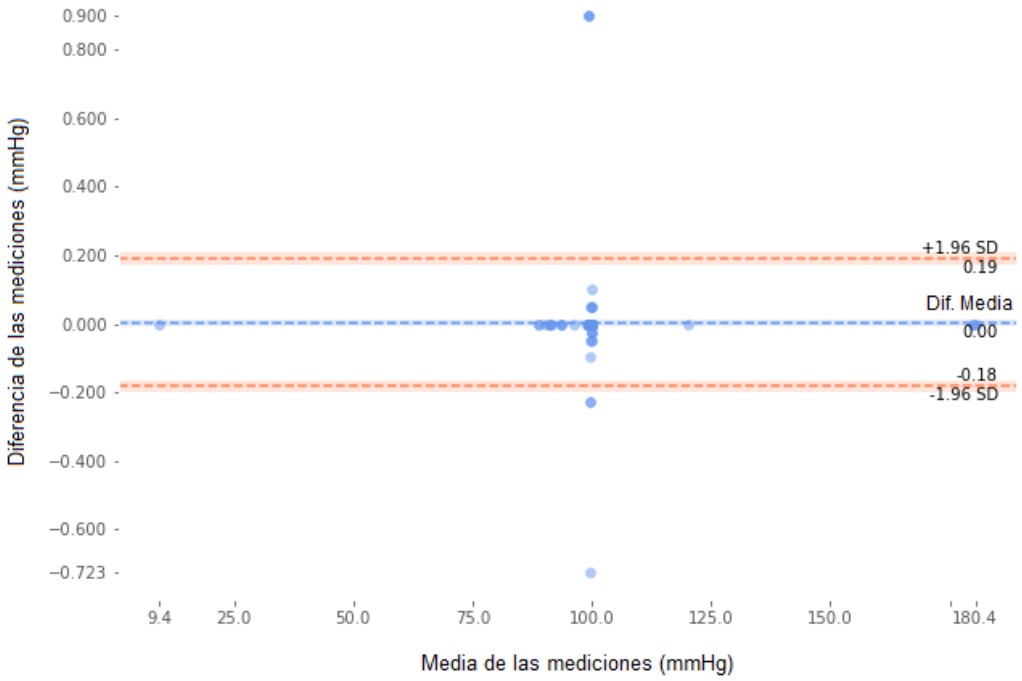

Figura 4.2: Gráfico de Bland-Altman para la estimación de la PAS.

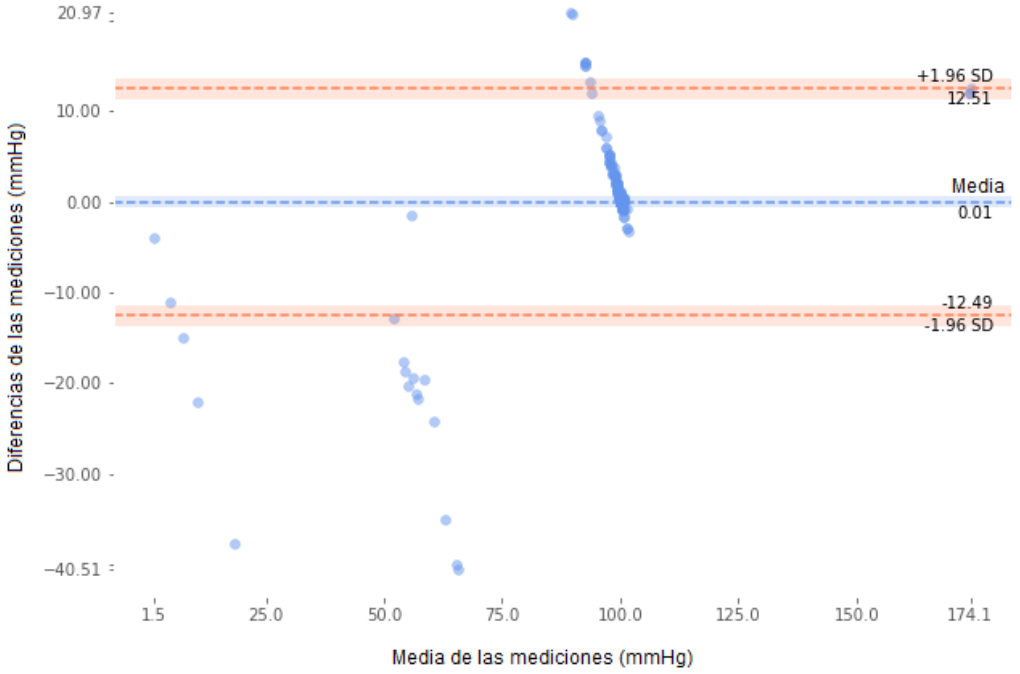

Figura 4.3: Gráfico de Bland-Altman para la estimación de la PAD.

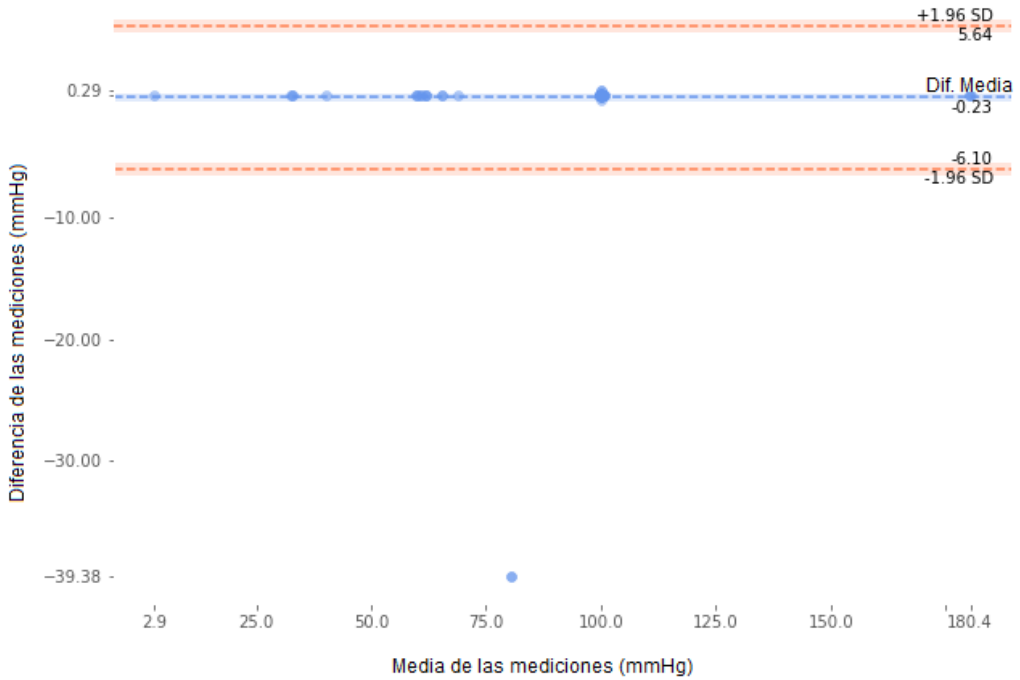

Figura 4.4: Gráfico de Bland-Altman para la estimación de la PAM.

CAPÍTULO 5

CONCLUSIONES Y TRABAJOS FUTUROS

En esta tesis se propuso la reducción de la dimensionalidad del conjunto de datos mediante las exploraciones de las relaciones multivariadas lineales y no lineales de las señales de fotopletismografía a través del análisis de componentes principales.

Y las conclusiones son:

- Las matrices de las correlaciones de Pearson y Spearman presentaron correlaciones positivas y en gran parte de moderadas a altas. Es mínima la diferencia que se presentan entre las correlaciones lineales y no lineales del conjunto de datos.

- La matriz de correlación de Spearman presenta levemente una mayor varianza acumulada desde el primer componente principal con respecto a los obtenidos con la matriz de correlación de Pearson.

- Para el entrenamiento de los modelos predictivos se seleccionaron los primeros 2, 3, 5 y 8 componentes principales, los cuales retenían el 70 %, el 80 %, el 90 % y el 96 % de varianzas acumuladas con respecto al conjunto de datos original.

- Los modelos predictivos con mejores rendimientos fueron los que se entrenaron con los componentes principales reducidos a través del análisis de componentes principales mediante la descomposición de la matriz de correlación de Spearman. Los modelos predictivos entrenados con 5 componentes principales obtuvieron mejores rendimientos para la estimación de la PA con respecto al entrenamiento con los demás componentes principales.

Para trabajos futuros se recomienda seleccionar diferentes tipos de señales de fotopletismografía, y realizar un análisis comparativo con otras técnicas de reducción de atributos.

REFERENCIAS

[1] Organización Panamericana de la Salud. *Especificaciones técnicas de la OMS para dispositivos automáticos de medición de la presión arterial no invasivos y con brazalete*. Inf. téc. 2020.

[2] U.S. National Library of Medicine. *Vital signs*. 2021.

[3] World Health Organization. *Hypertension*. 2021.

[4] Andrew Reisner, Phillip A Shaltis, Devin McCombie, H Harry Asada, David S Warner y Mark A Warner. «Utility of the photoplethysmogram in circulatory monitoring». En: *The Journal of the American Society of Anesthesiologists* 108.5 (2008), págs. 950-958.

[5] Seyedeh Somayyeh Mousavi, Mohammad Firouzmand, Mostafa Charmi, Mohammad Hemmati, Maryam Moghadam y Yadollah Ghorbani. «Blood pressure estimation from appropriate and inappropriate PPG signals using A whole-based method». En: *Biomedical Signal Processing and Control* 47 (2019), págs. 196-206.

[6] L Breiman, JH Friedman, R Olshen y CJ Stone. «Classification and Regression Trees». En: (1984).

[7] Vladimir N Vapnik. *The nature of statistical learning theory*. 1995.

[8] Yoav Freund y Robert E Schapire. «A decision-theoretic generalization of on-line learning and an application to boosting». En: *Journal of computer and system sciences* 55.1 (1997), págs. 119-139.

[9] Leo Breiman. «Random forests». En: *Machine learning* 45.1 (2001), págs. 5-32.

[10] Ludi Wang, Wei Zhou, Ying Xing y Xiaoguang Zhou. «A novel neural network model for blood pressure estimation using photoplethesmography without electrocardiogram». En: *Journal of healthcare engineering* 2018 (2018).

[11] Warren S McCulloch y Walter Pitts. «A logical calculus of the ideas immanent in nervous activity». En: *The bulletin of mathematical biophysics* 5.4 (1943), págs. 115-133.

[12] Syed Ghufran Khalid, Jufen Zhang, Fei Chen y Dingchang Zheng. «Blood pressure estimation using photoplethysmography only: comparison between different machine learning approaches». En: *Journal of healthcare engineering* 2018 (2018).

[13] Naoki Watanabe, Yasuko K Bando, Taiji Kawachi, Hiroshi Yamakita, Kouki Futatsuyama, Yoshikazu Honda, Hisae Yasui, Kazuyuki Nishimura, Takahiro Kamihara, Takahiro Okumura y col. «Development and validation of a novel cuff-less blood pressure monitoring device». En: *Basic to Translational Science* 2.6 (2017), págs. 631-642.

[14] T. Nepusz, A. Petróczi, L. Négyessy y F. Bazsó. «Fuzzy communities and the concept of bridgeness in complex networks». En: *Phys. Rev. E.* 77 (2008), pág. 016107.

[15] Max Kuhn, Kjell Johnson y col. *Applied predictive modeling.* Vol. 26. Springer, 2013.

[16] Ian T Jolliffe y Jorge Cadima. «Principal component analysis: a review and recent developments». En: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374.2065 (2016), pág. 20150202.

[17] Karl Pearson. «LIII. On lines and planes of closest fit to systems of points in space». En: *The London, Edinburgh, and Dublin philosophical magazine and journal of science* 2.11 (1901), págs. 559-572.

[18] Harold Hotelling. «Analysis of a complex of statistical variables into principal components.» En: *Journal of educational psychology* 24.6 (1933), pág. 417.

[19] Richard Arnold Johnson, Dean W Wichern y col. *Applied multivariate statistical analysis.* Vol. 6. Pearson London, UK: 2014.

[20] Felix Scholkmann, Jens Boss y Martin Wolf. «An efficient algorithm for automatic peak detection in noisy periodic and quasi-periodic signals». En: *Algorithms* 5.4 (2012), págs. 588-603.

[21] MJMMJ Mukaka. «Statistics corner: a guide to appropriate use of correlation in medical research». En: *Malawi Medical Journal* 24.3 (2012), págs. 69-71.

[22] Gianluca Egidi, Magda Edwards, Sirio Cividino, Filippo Gambella y Luca Salvati. «Exploring non-linear relationships among redundant variables through non-parametric principal component analysis: An empirical analysis with land-use data». En: *Regional Statistics* 11.1 (2021), págs. 25-41.

[23] Nian Shong Chok. «Pearson's versus Spearman's and Kendall's correlation coefficients for continuous data». Tesis doct. University of Pittsburgh, 2010.

[24] A. Goldberger, L. Amaral, L. Glass, J. Hausdorff, P. C. Ivanov, R. Mark y H. E Stanley. *PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals.* 2000.

[25] Aurélien Géron. «Hands-on machine learning with scikit-learn and tensorflow: Concepts». En: *Tools, and Techniques to build intelligent systems* (2017).

[26] Liliane Bel, Denis Allard, Jean-Marie Laurent, Rachid Cheddadi y Avner Bar-Hen. «CART algorithm for spatial data: Application to environmental and ecological data». En: *Computational Statistics & Data Analysis* 53.8 (2009), págs. 3082-3093.

[27] Guido Van Rossum y Fred L Drake Jr. *Python tutorial*. Vol. 620. Centrum voor Wiskunde en Informatica Amsterdam, 1995.

[28] George S Stergiou, Bruce Alpert, Stephan Mieke, Roland Asmar, Neil Atkins, Siegfried Eckert, Gerhard Frick, Bruce Friedman, Thomas Graßl, Tsutomu Ichikawa y col. «A universal standard for the validation of blood pressure measuring devices: Association for the Advancement of Medical Instrumentation/European Society of Hypertension/International Organization for Standardization (AAMI/ESH/ISO) Collaboration Statement». En: *Hypertension* 71.3 (2018), págs. 368-374.

[29] J Martin Bland y Douglas G Altman. «Measuring agreement in method comparison studies». En: *Statistical methods in medical research* 8.2 (1999), págs. 135-160.

[30] Eoin O'brien, Bernard Waeber, Gianfranco Parati, Jan Staessen y Martin G Myers. «Blood pressure measuring devices: recommendations of the European Society of Hypertension». En: *Bmj* 322.7285 (2001), págs. 531-536.

APÉNDICE

A.1. Conjunto de Datos - Señales de Fotopletismografía

A continuación se presentan algunas de las señales de fotopletismografía (PPG) obtenidas desde la Base de Datos MIMIC [24].

Las señales de PPG que presentaban artefactos incorregibles fueron descartadas para el proceso de estimación de la PA.

De las cinco figuras presentadas a continuación, las Figuras A.4 y A.5 fueron descartadas. Estas figuras son sólo algunos de los tipos de señales que se analizaron en la primera etapa del proceso para la estimación de la PA.

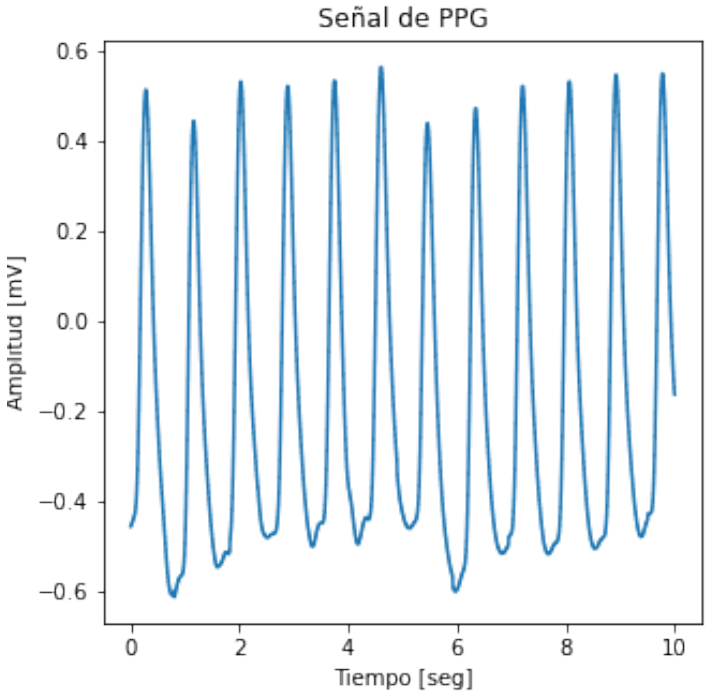

Figura A.1: Segmento de la señal de PPG del registro 0001.

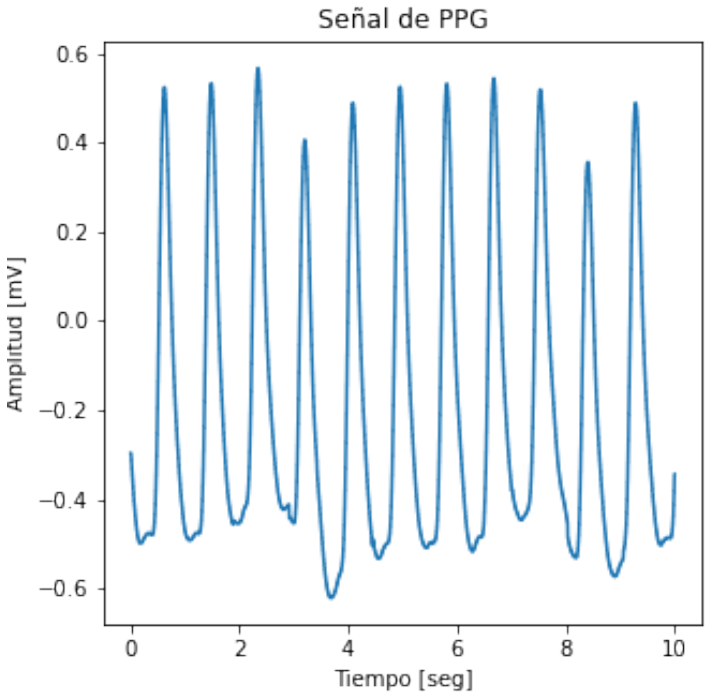

Figura A.2: Segmento de la señal de PPG del registro 0002.

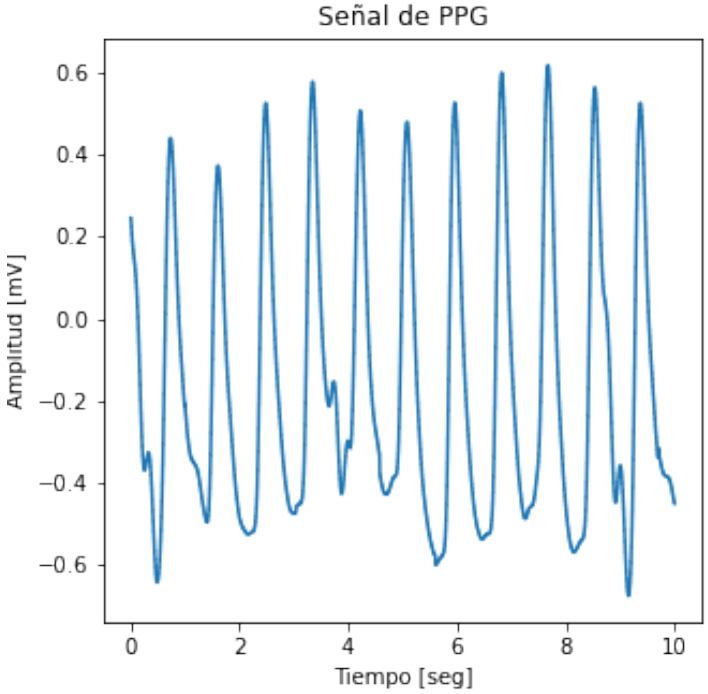

Figura A.3: Segmento de la señal de PPG del registro 0004.

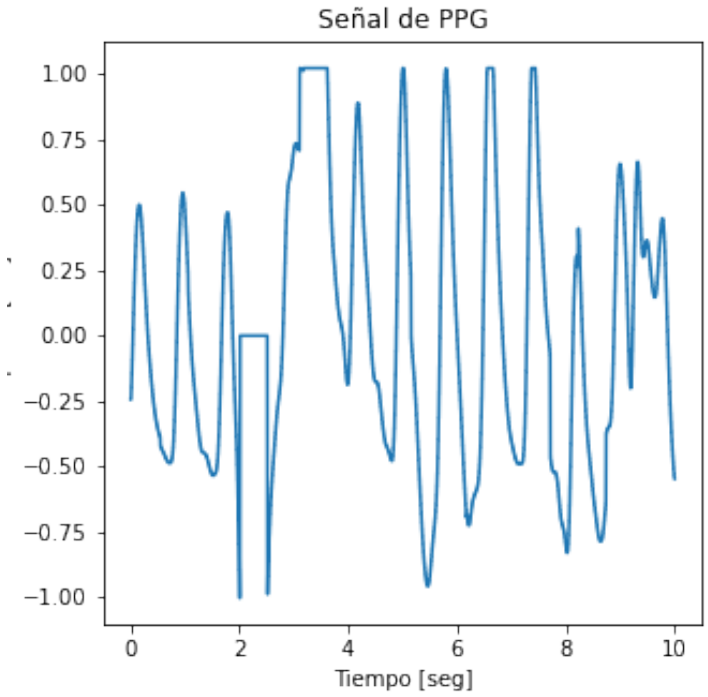

Figura A.4: Segmento de la señal de PPG del registro 0005.

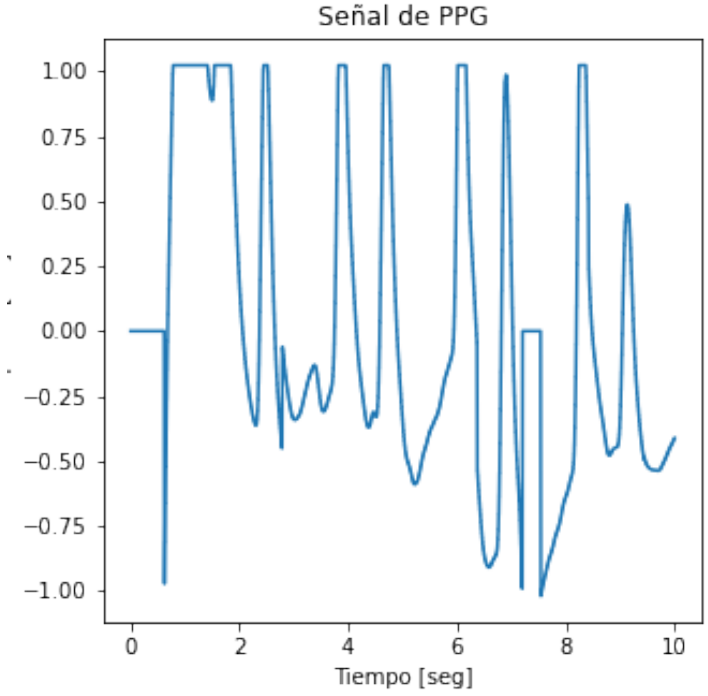

Figura A.5: Segmento de la señal de PPG del registro 0048.

A.2. Publicación y Difusión.

A continuación se presentan los resultados obtenidos de un trabajo preliminar a esta tesis. Dicho trabajo fue seleccionado y presentado en la Conferencia Ibero Americana de Computación Aplicada (CIACA) en el mes de noviembre del 2021.

A.2.1. Información del Artículo Presentado

Título: Reducción de Características Utilizando Correlación de Spearman y Análisis de Componentes Principales para la Estimación de la Presión Arterial.

Resumen: Determinar el número óptimo de predictores que deben incluirse en un modelo de predicción es una de las cuestiones críticas a medida que el conjunto de datos va incrementando. Además, algunos modelos pueden verse afectados por predictores no informativos. Aplicar la técnica de reducción de características apropiada permitirá reducir la cantidad de variables numéricas de entrada no informativa o redundante, y mejorar el rendimiento de los modelos de predicción. El propósito del trabajo es la reducción de características utilizando Correlación de Spearman y Análisis de Componentes Principales, aplicadas con algoritmos de aprendizaje automático para la estimación de la presión arterial utilizando señales de fotopletismografía. La evaluación de los modelos predictivos aplicando las dos técnicas de reducción de características con diferentes enfoques, se realizó comparando las métricas de regresión ME±SD, MAE, MSE y RMSE para la estimación de la presión arterial sistólica (PAS), la presión arterial diastólica (PAD) y la presión arterial media (PAM). Los resultados comparativos demostraron que la técnica de reducción de características aplicando análisis de componentes principales como paso previo al entrenamiento de los modelos predictivos, permite obtener mejores desempeños de predicción para la PAS (-0,14±1,66 mmHg), la PAD (0,40±6,40 mmHg) y la PAM (0,37±5,64 mmHg). Las técnicas de reducción de características en conjunto con los modelos predictores propuestos, presentan valores de estimación que se posicionan en el rango de precisión de los estándares AAMI y BHS.

A.2.2. Resultados

En las Figuras A.6, A.7 y A.8 se presentan los resultados de las métricas de evaluación de los modelos predictivos para las estimaciones de la PAS, la PAD y la PAM, teniendo en cuenta la cantidad de predictores seleccionados por las diferentes técnicas de reducción de la dimensionalidad (TRD) aplicadas. Se implementaron varias funciones de pérdida, puntuación y utilidad para medir el rendimiento de los modelos predictivos de regresión. Las métricas de regresión aplicadas fueron: error medio (ME) ± desviación estándar (SD), error absoluto medio (MAE), error cuadrático medio (MSE), y la raíz del error cuadrático medio (RMSE).

El número de predictores seleccionados como relevantes por cada una de las TRD aplicadas, se pueden observar en la columna números (Nº) de predictores de las Figuras A.6, A.7 y A.8. Para determinar los modelos predictivos con mejores desempeños para las estimaciones de la PAS, la PAD y la PAM se tuvieron en cuenta las siguientes métricas de evaluación: ME±SD, MAE, MSE y RMSE. Se realizó un ranking con los 10 valores mínimos de ME±SD y MAE, a partir del mismo, se seleccionó el modelo que presentaba menor MSE y RMSE. Cabe resaltar que estás últimas métricas, castigan los errores grandes de predicción entre el valor estimado y el valor esperado.

ML	TSC/TRD	Corr.	Varianza Acumulada	N°	ME±SD	MAE	MSE	RMSE
	No aplica	-	-	618	0,65 ± 0,84	0,65	44,13	6,64
DTR	Filtro (CSp)	>=0,9	-	17	0,64 ± 7,71	0,81	59,81	7,73
		>=0,7	-	60	0,61 ± 6,28	0,66	39,83	6,31
		>=0,5	-	214	-0,23 ± 4,94	0,36	24,41	4,94
	PCA	-	95%	8	0,47 ± 6,14	0,48	37,94	6,16
		-	99%	15	0 ± 8,79	1,09	77,27	8,79
SVR	No aplica	-	-	618	0,21 ± 9,13	1,38	83,40	9,13
	Filtro (CSp)	>=0,9	-	17	0,21 ± 9,13	1,40	83,39	9,13
		>=0,7	-	60	0,20 ± 9,13	1,40	83,36	9,13
		>=0,5	-	214	0,20 ± 9,13	1,39	83,38	9,13
	PCA	-	95%	8	0,20 ± 9,13	1,38	83,39	9,13
		-	99%	15	0,20 ± 9,13	1,37	83,39	9,13
AdaBoostR	No aplica	-	-	618	1,20 ± 0,36	1,20	4,81	2,19
	Filtro (CSp)	>=0,9	-	17	-1,45 ± 12,65	8,48	162,18	12,73
		>=0,7	-	60	0,82 ± 10,16	6,46	103,90	10,19
		>=0,5	-	214	0,04 ± 4,56	1,54	20,83	4,56
	PCA	-	95%	8	-0,25 ± 3,26	1,10	10,66	3,27
		-	**99%**	**15**	**-0,14 ± 1,66**	**0,61**	**2,76**	**1,66**
RFR	No aplica	-	-	618	0,80 ± 0,53	0,80	13,69	3,70
	Filtro (CSp)	>=0,9	-	17	-0,31 ± 7,97	1,69	63,65	7,98
		>=0,7	-	60	-0,05 ± 6,51	1,46	42,42	6,51
		>=0,5	-	214	0,01 ± 3,48	0,83	12,08	3,48
	PCA	-	95%	8	0 ± 4,34	0,93	18,86	4,34
		-	99%	15	0,02 ± 5,57	1,19	31,05	5,57

Figura A.6: Métricas de evaluación de los modelos predictivos para la estimación de la PAS.

En la Figura A.6 se observa que el modelo AdaBoostR con la TRD (PCA) con 15 CPs, fue el modelo que presentaba mejor desempeño para la estimación de la PAS. En la Figura A.7 se observa que el modelo Random Forest Regression con la TRD (PCA) con 8 CPs, fue el modelo que presentaba mejor desempeño para la estimación de la PAD. En la Figura A.8 se observa que el modelo Random Forest Regression con la TRD (PCA) con 8 CPs, fue el modelo

ML	TSC/TRD	Corr.	Varianza Acumulada	N°	ME±SD	MAE	MSE	RMSE
	No aplica	-	-	**618**	1,88 ± 0,74	1,88	128,74	11,35
DTR	Filtro (CSp)	>=0,9	-	17	-0,28 ± 10,58	1,90	112,09	10,59
		>=0,7	-	60	0,17 ± 8,61	1,42	74,19	8,61
		>=0,5	-	214	0,94 ± 8,89	1,27	79,97	8,94
	PCA	-	95%	8	0,89 ±12,77	1,81	163,97	12,80
		-	99%	15	0,27 ±14,99	2,75	224,86	15
SVR	No aplica	-	-	618	-2,72 ± 17,41	4,27	310,25	17,61
	Filtro (CSp)	>=0,9	-	17	-2,72 ± 17,41	4,29	310,39	17,62
		>=0,7	-	60	-2,72 ± 17,40	4,29	310,22	17,61
		>=0,5	-	214	-2,72 ± 17,40	4,28	310,26	17,61
	PCA	-	95%	8	-2,73 ± 17,40	4.26	310,24	17,61
		-	99%	15	-2,73 ±17,40	4,26	310,04	17,61
AdaBoostR	No aplica	-	-	618	13,83 ± 2,27	13,83	246,18	15,69
	Filtro (CSp)	>=0,9	-	17	15,47 ± 17,91	19,99	560,24	23,67
		>=0,7	-	60	6,77 ± 23,15	20,84	581,88	24,12
		>=0,5	-	214	7,46 ± 22,64	22,37	568,08	23,83
	PCA	-	95%	8	10,96 ± 13,64	15,14	306,11	17,50
		-	99%	15	9,64 ± 13,36	13,08	271,20	16,47
RFR	No aplica	-	-	618	2,55 ± 0,50	2,55	49,75	7,05
	Filtro (CSp)	>=0,9	-	17	-0,04 ± 11,32	3,60	127,93	11,31
		>=0,7	-	60	0,22 ± 9,17	3,05	84,16	9,17
		>=0,5	-	214	-0,15 ± 7,67	2,58	58,84	7,67
	PCA	-	**95%**	**8**	**0,40 ± 6,40**	**2,39**	**41,13**	**6,41**
		-	99%	15	0,30 ± 7,12	2,81	50,74	7,12

Figura A.7: Métricas de evaluación de los modelos predictivos para la estimación de la PAD.

ML	TSC/TRD	Corr.	Varianza Acumulada	N°	ME±SD	MAE	MSE	RMSE
	No aplica	-	-	618	1,40 ± 0,53	1,40	76,68	8,76
DTR	Filtro (CSp)	>=0,9	-	17	0,20 ± 7,92	1,21	62,85	7,93
		>=0,7	-	60	-0,14 ± 6,80	1,11	46,27	6,80
		>=0,5	-	214	-0,25 ± 6,30	0,89	39,80	6,31
	PCA	-	95%	8	-0,41 ± 8,91	0,66	24,29	4,93
		-	99%	15	-0,77 ± 12,57	2,54	158,61	12,59
SVR	No aplica	-	-	618	-1,75 ± 13,58	3,24	187,43	13,69
	Filtro (CSp)	>=0,9	-	17	-1,75 ± 13,58	3,25	187,50	13,69
		>=0,7	-	60	-1,75 ± 13,58	3,25	187,40	13,69
		>=0,5	-	214	-1,75 ± 13,58	3,25	187,49	13,69
	PCA	-	95%	8	-1.74 ± 13,58	3.23	187,36	13,69
		-	99%	15	-1,75 ± 13,57	3,23	187,28	13,69
AdaBoostR	No aplica	-	-	618	12,28 ± 2,86	12,28	181,09	13,46
	Filtro (CSp)	>=0,9	-	17	9,77 ± 12,86	12,94	260,86	16,15
		>=0,7	-	60	3,34 ± 18,58	16,33	356,34	18,88
		>=0,5	-	214	6,83 ± 15,98	15,52	301,92	17,38
	PCA	-	95%	8	10,28 ± 10,48	12,72	215,42	14,68
		-	99%	15	9,26 ± 10,83	11,66	203,03	14,25
RFR	No aplica	-	-	618	1,93 ± 0,47	1,93	29,50	5,43
	Filtro (CSp)	>=0,9	-	17	0 ± 9,78	2,85	95,70	9,78
		>=0,7	-	60	0,18 ± 7,97	2,53	63,54	7,97
		>=0,5	-	214	-0,04 ± 5,97	1,95	35,65	5,97
	PCA	-	**95%**	**8**	**0,37 ± 5,64**	**1,92**	**31,93**	**5,65**
		-	99%	15	0,21 ± 6,73	2,36	45,36	6,73

Figura A.8: Métricas de evaluación de los modelos predictivos para la estimación de la PAM.

que presentaba mejor desempeño para la estimación de la PAM. Los resultados comparativos demostraron que la TRD aplicando PCA como paso previo al entrenamiento de los modelos predictivos, permitieron obtener mejores desempeños de estimaciones.

En las Figuras A.9, A.10 y A.11, se presentan tres gráficos de Bland-Altman correspondientes a la diferencia del valor estimado con respecto al valor esperado para las estimaciones de la PAS, la PAD y la PAM. Con este método gráfico, cuantificamos la diferencia de ambas mediciones, y un intervalo de confianza, que se espera que incluyan el 95 % de la diferencia de ambas mediciones. Como se pueden observar en los tres gráficos, las diferencias de mediciones para cada una de las estimaciones, se encuentran en su mayoría dentro del intervalo de confianza con una dispersión mínima con respecto a la misma.

Para validar la exactitud clínica de los dispositivos de medición de la PA no invasiva, se disponen de protocolos estandarizados. Los parámetros principales considerados en los protocolos de validación de la AAMI y la Sociedad Británica de Hipertensión (BHS, por sus siglas en inglés) señalan lo siguiente: el número de participantes debe ser de 85 o más (ambos) [28], [30]. Para la AAMI, la diferencia media global es de ± 5mmHg, y la desviación estándar no debe superar los 8mmHg [28]. Para la BHS, la proporción de mediciones de la PA (referencia de la prueba) total se dividen por grados: dentro de un margen de 5mmHg (Grado A), 10mmHg (Grado B) y 15mmHg (Grado C) [30]. En el presente trabajo se utilizaron los registros de al menos 90 participantes según la Base de Datos MIMIC [24].

Los modelos predictivos con mejores desempeños para las estimaciones de la PAS, la PAD y la PAM se pueden observar en la Tabla A.1. Para los tres casos, la TRD no supervisada (PCA) logró que los modelos predictivos seleccionados presentaran buenos desempeños de estimaciones. Además, se pueden observar que los modelos predictivos con las técnicas seleccionadas para cada estimación, se encuentran dentro del rango del estándar de la AAMI. En tanto que para el estándar de la BHS, la estimación de la PAS se clasifica en el grado A y las estimaciones de la PAD y la PAM se clasifican en el grado B.

Tabla A.1: Modelos predictivos con mejores rendimientos para las estimaciones de la PAS, la PAD y la PAM con respecto al estándar AAMI.

PA			
Resultados	Modelo predictivo y TRD	N°	ME±STD
PAS	AdaBoostR con PCA	15	-0,14±1,66
PAD	RFR con PCA	8	0,40±6,40
PAM	RFR con PCA	8	0,37±5,64
	AAMI (STANDARD)		$(\leq 5) \pm (\leq 8)$

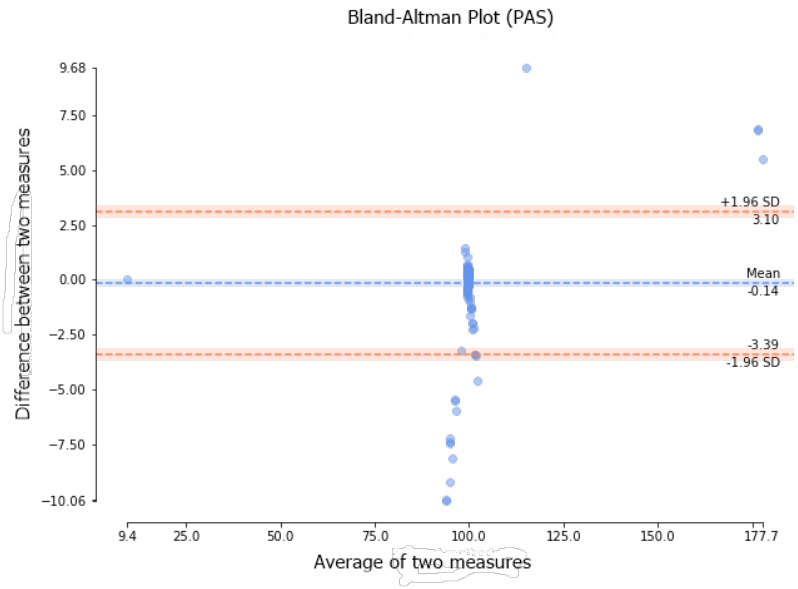

Figura A.9: Gráfico de Bland-Altman para la estimación de la PAS.

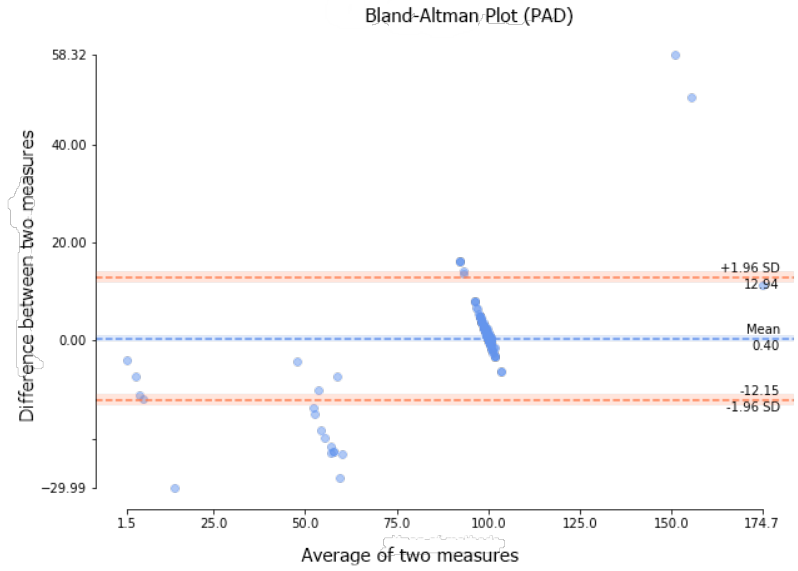

Figura A.10: Gráfico de Bland-Altman para la estimación de la PAD.

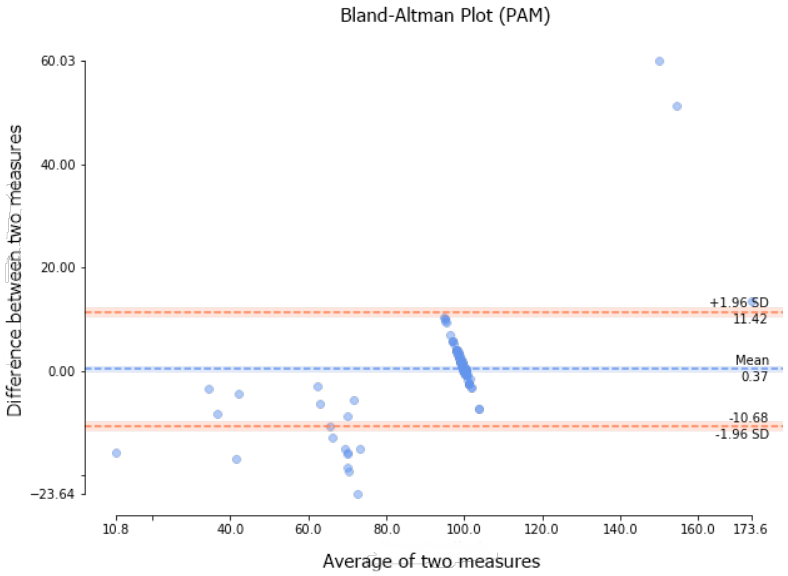

Figura A.11: Gráfico de Bland-Altman para la estimación de la PAM.

A.2.3. Discusión y Conclusión

En este estudio se analizaron las técnicas de reducción de características utilizando correlación de Spearman y análisis de componentes principales, aplicadas con algoritmos de aprendizaje automático para la estimación de la presión arterial utilizando señales de fotopletismografía. Los resultados comparativos de los experimentos demuestran que el modelo AdaBoostR con la técnica de reducción de dimensionalidad (PCA) con 15 componentes principales, es el modelo que presenta mejor desempeño para la estimación de la PAS. Y el modelo Random Forest Regression con la técnica de reducción de dimensionalidad (PCA) con 8 componentes principales, es el modelo que presenta mejor desempeño para las estimaciones de la PAD y la PAM. Los resultados comparativos demuestran que la técnica de reducción de características aplicando análisis de componentes principales (PCA) como paso previo al entrenamiento de los modelos predictivos, permiten obtener mejores desempeños de estimaciones para la PAS (-0,14±1,66 mmHg), la PAD (0,40±6,40 mmHg) y la PAM (0,37±5,64 mmHg). Las técnicas de reducción de características en conjunto con los modelos predictores propuestos, presentan valores de estimaciones que se posicionan en el rango de precisión del estándar de la AAMI. Y para el estándar de la BHS, la estimación de la PAS se clasifica en el grado A y las estimaciones de la PAD y la PAM se clasifican en el grado B. Para trabajos futuros se recomienda caracterizar y clasificar cada uno de los componentes principales, y aplicar otras técnicas de selección de atributos con algoritmos de aprendizaje automático.

Y finalmente en la Figura A.12, se presenta el certificado de la Mención Honorífica recibida por la presentación del trabajo en la Conferencia Ibero-Americana de Computación Aplicada 2021.

Certificado

El artículo titulado

REDUCCIÓN DE CARACTERÍSTICAS UTILIZANDO CORRELACIÓN DE SPEARMAN Y ANÁLISIS DE COMPONENTES PRINCIPALES PARA LA ESTIMACIÓN DE LA PRESIÓN ARTERIAL

de

Carolina Elizabeth Villegas Colmán[1], Cynthia Emilia Villalba Cardozo[1], José Luis Vázquez Noguera[1] y Miguel García Torres[2]
[1]Facultad Politécnica – Universidad Nacional de Asunción, Paraguay
[2]Universidad Pablo de Olavide, España

Recibió una

Mención Honorífica

en la

Conferência Ibero-Americana Computación Aplicada 2021

El Comité de la Conferencia, tomando en consideración el resultado del proceso ciego de revisión, considera este artículo de la máxima calidad.

Paula Miranda
Program Chair

iadis
international association for development of the information society

Figura A.12: Certificado de la Mención Honorífica recibida en la Conferencia Ibero-Americana de Computación Aplicada 2021.