LOS MODELOS DE CLASIFICACIÓN EMPLEANDO PCA PARA PREDECIR ENFERMEDADES CARDÍACAS

Yuliza Angela Huanca Mamani

Universidad Mayor de San Andrés, Facultad de Ciencias Puras y Naturales, Carrera de Informática

[yhuancam@fcpn.edu.bo](mailto:yhuancam@fcpn.edu.bo)

|  |  |
| --- | --- |
| **Resumen** | **Abstract** |
| La predicción de enfermedades cardiacas coadyuva al personal médico a tomar decisiones con más precisión con respecto a la salud de los pacientes. El uso de aprendizaje automático es una solución para reducir y comprender los síntomas relacionados con las enfermedades del corazón. El objetivo del presente trabajo es la propuesta de un método de reducción de dimensionalidad y búsqueda de características de enfermedades del corazón mediante la aplicación de una técnica de selección de características. La información utilizada para este análisis.  ***Palabras Clave:*** predicción, enfermedades cardiacas, decisiones, precisión, salud, pacientes, aprendizaje automático, reducir, comprender, síntomas, enfermedades del corazón, método, reducción de dimensionalidad, búsqueda de características, técnica de selección de características, análisis, información. | The prediction of heart diseases assists medical personnel in making more accurate decisions regarding the health of patients. The use of machine learning is a solution to reduce and understand symptoms related to heart diseases. The objective of this study is to propose a dimensionality reduction method and feature search for heart diseases through the application of a feature selection technique. The information used for this analysis.  ***Keywords:*** prediction, heart disease, decisions, accuracy, health, patients, machine learning, reduce, understand, symptoms, heart disease, method, dimensionality reduction, feature search, feature selection technique, analysis, information. |

# 1. Introducción

# Las enfermedades cardiacas son responsables de una proporción significativa de las muertes a nivel mundial. En el año 2019, representaron aproximadamente un tercio de todas las muertes. Durante este período, los casos de enfermedades cardiacas experimentaron un aumento notable, casi duplicándose desde 271 millones en 1990 hasta 523 millones en 2019. Asimismo, el número de muertes relacionadas con estas enfermedades aumentó de 12,1 millones a 18,6 millones.

# Identificar las causas de las enfermedades cardiacas y prevenirlas se ha convertido en un desafío fundamental. El razonamiento de diagnóstico médico se está volviendo cada vez más popular para abordar esta problemática.

# La predicción de enfermedades cardíacas es considerada uno de los temas más importantes en el ámbito de la salud. Sin embargo, predecir la aparición de una enfermedad cardíaca es una tarea compleja que requiere un alto nivel de experiencia y conocimiento para obtener resultados precisos. Según una encuesta realizada por la Organización Mundial de la Salud (OMS), los profesionales médicos logran predecir correctamente las enfermedades cardíacas con una precisión de apenas el 67%.

El Análisis de Componentes Principales (PCA) se utiliza para crear nuevos componentes que capturan la información más relevante de las características originales al preservar una alta varianza. En el ámbito de la salud, se han llevado a cabo diversos estudios que emplean PCA como técnica de extracción de características para la clasificación.

Los conceptos matemáticos que se aplican en el PCA: eigenvectors y eigenvalues. Se trata simplemente de una descripción intuitiva con la única finalidad de facilitar el entendimiento del cálculo de componentes principales.

Los eigenvectors son un caso particular de multiplicación entre una matriz y un vector.

El vector resultante de la multiplicación es un múltiplo entero del vector original. Los eigenvectors de una matriz son todos aquellos vectores que, al multiplicarlos por dicha matriz, resultan en el mismo vector o en un múltiplo entero del mismo. Los eigenvectors tienen una serie de propiedades matemáticas específicas:

Los eigenvectors solo existen para matrices cuadradas y no para todas. En el caso de que una matriz n x n tenga eigenvectors, el número de ellos es n.

Si se escala un eigenvector antes de multiplicarlo por la matriz, se obtiene un múltiplo del mismo eigenvector. Esto se debe a que, si se escala un vector multiplicándolo por cierta cantidad, lo único que se consigue es cambiar su longitud, pero la dirección es la misma.

Todos los eigenvectors de una matriz son perpendiculares (ortogonales) entre ellos, independientemente de las dimensiones que tengan.

Dada la propiedad de que multiplicar un eigenvector solo cambia su longitud, pero no su naturaleza de eigenvector, es frecuente escalarlos de tal forma que su longitud sea 1.

Eigenvalue, cuando se multiplica una matriz por alguno de sus eigenvectors se obtiene un múltiplo del vector original, es decir, el resultado es ese mismo vector multiplicado por un número. Al valor por el que se multiplica el eigenvector resultante se le conoce como eigenvalue. A todo eigenvector le corresponde un eigenvalue y viceversa.

En el método PCA, cada una de las componentes se corresponde con un eigenvector, y el orden de componente se establece por orden decreciente de eigenvalue. Así pues, la primera componente es el eigenvector con el eigenvalue asociado más alto.

La interpretación geométrica de las componentes principales, suponiendo un conjunto de observaciones para las que se dispone de dos variables . El vector que define la primera componente principal sigue la dirección en la que las observaciones varían más. La proyección de cada observación sobre esa dirección equivale al valor de la primera componente para dicha observación.

La segunda componente sigue la segunda dirección en la que los datos muestran mayor varianza y no esta correlacionada con la primera componente. La condición de no correlación entre componentes principales equivale a decir que sus direcciones son perpendiculares/ortogonales.

Cada componente principal se obtiene por combinación lineal de las variables originales. Se pueden entender como nuevas variables obtenidas al combinar de una determinada forma las variables originales. La primera componente principal de un grupo de variables es la combinación lineal normalizada de dichas variables que tiene mayor varianza.

Que la combinacion lineal sea normalizada implica que:

Los terminos reciben en el nombre de loadings y son los que definen a la componente es el loading de la variable de la primera componente principal.

Dado un set de datos X

con n observaciones y p variables, el proceso a seguir para calcular la primera componente principal es:

Centralización de las variables: se resta a cada valor la media de la variable a la que pertenece. Con esto se consigue que todas las variables tengan media cero.

Se resuelve un problema de optimización para encontrar el valor de los loadings con los que se maximiza la varianza. Una forma de resolver esta optimización es mediante el cálculo de eigenvector-eigenvalue de la matriz de covarianzas.

Una vez calculada la primera componente se calcula la segunda repitiendo el mismo proceso, pero añadiendo la condición de que la combinación lineal no pude estar correlacionada con la primera componente. Esto equivale a decir que y .

tienen que ser perpendiculares. EL proceso se repite de forma iterativa hasta calcular todas las posibles componentes (min (n-1, p)) o hasta que se decida detener el proceso. El orden de importancia de las componentes viene dado por la magnitud del eigenvalue asociado a cada eigenvector.

El proceso de PCA genera siempre las mismas componentes principales independientemente del software utilizado, es decir, el valor de los loadings resultantes es el mismo. La única diferencia que puede darse es que el signo de todos los loadings esté invertido. Esto es así porque el vector de loadings determina la dirección de la componente, y dicha dirección es la misma independientemente del signo (la componente sigue una línea que se extiende en ambas direcciones). Del mismo modo, el valor específico de las componentes obtenido para cada observación (principal component scores) es siempre el mismo, a excepción del signo.

Asumiendo que las variables se han normalizado para tener media cero, la varianza total presente en el set de datos se define como

Y la varianza explicada por la componente m es

Por lo tanto, la proporción de varianza explicada por la componente m viene dada por la ratio

Tanto la proporción de varianza explicada como la proporción de varianza explicada acumulada son dos valores de gran utilidad a la hora de decidir el número de componentes principales a utilizar en los análisis posteriores. Si se calculan todas las componentes principales de un set de datos, entonces, aunque transformada, se está almacenando toda la información presente en los datos originales. El sumatorio de la proporción de varianza explicada acumulada de todas las componentes es siempre 1.

Por lo general, dada una matriz de datos de dimensiones n x p, el número de componentes principales que se pueden calcular es como máximo de n-1 o p (el menor de los dos valores es el limitante). Sin embargo, siendo el objetivo del PCA reducir la dimensionalidad, suelen ser de interés utilizar el número mínimo de componentes que resultan suficientes para explicar los datos. No existe una respuesta o método único que permita identificar cual es el número óptimo de componentes principales a utilizar. Una forma de proceder muy extendida consiste en evaluar la proporción de varianza explicada acumulada y seleccionar el número de componentes mínimo a partir del cual el incremento deja de ser sustancial.

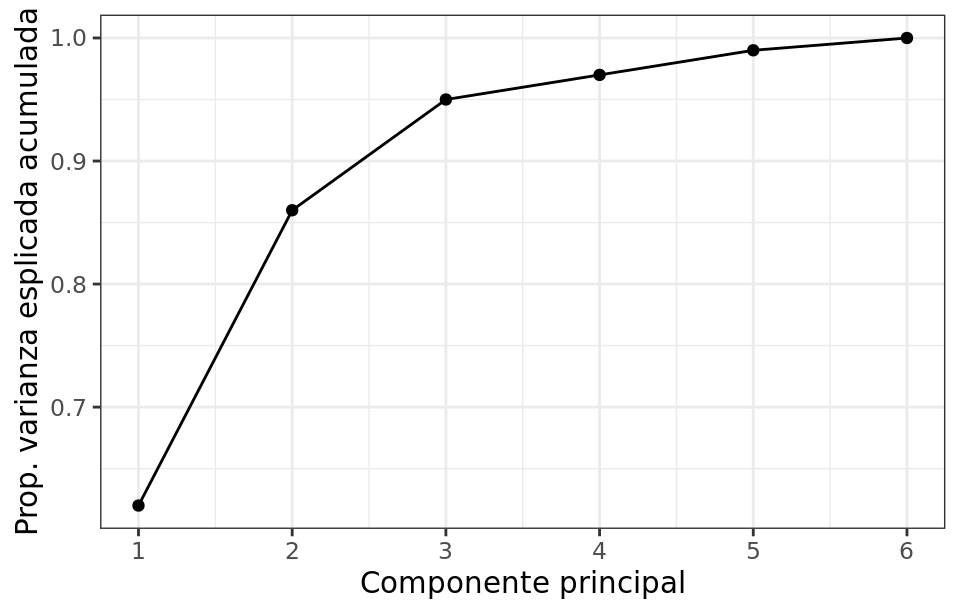


Imagen 1 Grafico de la varianza explicada vs componente principal

Los modelos de aprendizaje de clasificación con reducción de dimensionalidad tienen como objetivo principal aprender la mejor representación de características a partir de un conjunto de datos

El proceso de PCA se puede resumir en los siguientes pasos:

1. Estandarización de los datos: Antes de aplicar PCA, es necesario estandarizar los datos para asegurarse de que todas las variables tengan una media de 0 y una varianza de 1. Esto se hace para evitar que las variables con escalas mayores dominen a las variables con escalas menores.

2. Cálculo de la matriz de covarianza: Se calcula una matriz cuadrada que contiene las covarianzas entre todas las posibles parejas de variables en el conjunto de datos. Esta matriz ayuda a identificar las relaciones lineales entre las variables.

3. Cálculo de los autovalores y autovectores: Se determinan los autovalores y autovectores de la matriz de covarianza. Los autovectores representan las direcciones de los componentes principales, mientras que los autovalores indican la cantidad de varianza que cada componente principal puede explicar.

4. Selección de los componentes principales: Los autovalores se ordenan de mayor a menor, y se seleccionan los primeros k componentes principales, donde k es el número de componentes que se desea mantener. Esto se logra tomando los k autovectores correspondientes a los k autovalores más grandes.

5. Proyección de los datos: Finalmente, los datos originales se proyectan sobre los k componentes principales seleccionados, lo que crea un conjunto de datos de menor dimensión.

Estos pasos permiten reducir la dimensionalidad del conjunto de datos, preservando la información más relevante y explicativa en los componentes principales.

# 2. Materiales y Métodos

**2.1 Descripción del conjunto de datos**

El conjunto de datos utilizado en la presente investigación fue el conjunto de datos de enfermedades cardiacas del repositorio de Kaggle como se muestra en la imagen 1, teniendo una etiqueta llamada heart disease y 37 características independientes.

Heart disease especifica si un paciente se encuentra sano o con una enfermedad del corazón.

**2.2 Preprocesamiento de la información**

Los conjuntos de datos tenían variables comprendidas de 9 booleanas, 5 cadenas y 4 decimales como se puede observar en la imagen 2.

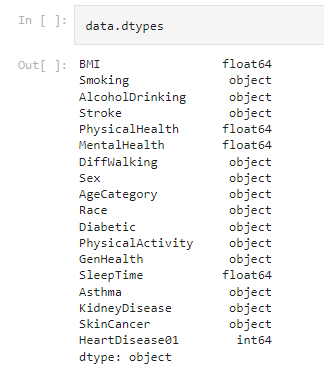


Imagen 2 Visualización de variables del conjunto de datos

Aplicamos la técnica dummy a las variables para que estas se encuentren en uniformidad numérica.

datos\_dummis = pd.get\_dummies(datos, drop\_first = True)

datos\_dummis.dtypes

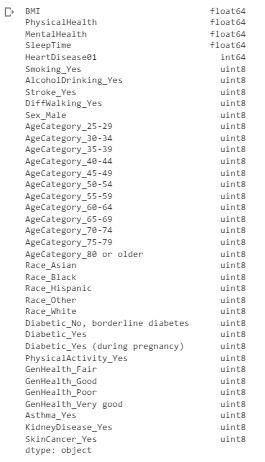


Imagen 3 Datos preprocesados del conjunto de datos

**2.3 Reducción de dimensionalidad**

Este scrip de Python realiza un Análisis de Componentes Principales (PCA) en un conjunto de datos almacenado en un archivo CSV llamado “heart\_2020\_cleaned.csv”. El objetivo de análisis es reducir la dimensionalidad del conjunto de datos, identificando y visualizando la varianza explicada por cada componente principal.

El código se puede dividir en las siguientes partes:

* Importación de librerías: Se importan las librerías necesarias, como pandas, numpy, StandardScaler, PCA y matplotlib.
* Lectura del archivo CSV: Se lee el archivo "plastic\_data.csv" y se guarda en el DataFrame 'df'. El índice de la columna 0 se utiliza como índice del DataFrame.
* Escalado de datos: Se crea una instancia de StandardScaler y se aplica a los datos para escalarlos.
* Aplicación de PCA: Se crea una instancia de PCA con un número determinado de componentes (3 en este caso) y se aplica a los datos escalados. Los componentes principales resultantes se almacenan en un nuevo DataFrame llamado 'principal\_df'.
* Visualización de la varianza explicada: Se crea un gráfico de barras que muestra la varianza explicada por cada uno de los componentes principales. También se imprime la varianza explicada y los componentes principales en sí.
* Gráfico de la varianza acumulativa explicada: Se calcula y grafica la varianza acumulativa explicada por los componentes principales. Se añade una línea horizontal en el 30% de la varianza acumulativa para identificar el número óptimo de componentes.
* Identificación de las variables originales en cada componente principal: Se crea un DataFrame llamado 'components\_df' que contiene las contribuciones de cada variable original en cada componente principal.
* Impresión de las contribuciones de las variables originales: Se imprime la tabla que muestra las contribuciones de cada variable original en cada componente principal.

scaler = StandardScaler()

scaled\_data = scaler.fit\_transform(datos\_dummis)

n\_components=11

pca = PCA(n\_components=n\_components)

principal\_components = pca.fit\_transform(scaled\_data)

components=[]

n=1

while n<=n\_components:

    components.append("PC"+str(n))

    n=n+1

principal\_df = pd.DataFrame(data=principal\_components, columns=components)

explained\_variance = pca.explained\_variance\_ratio\_

print("Varianza Explicada:", explained\_variance)



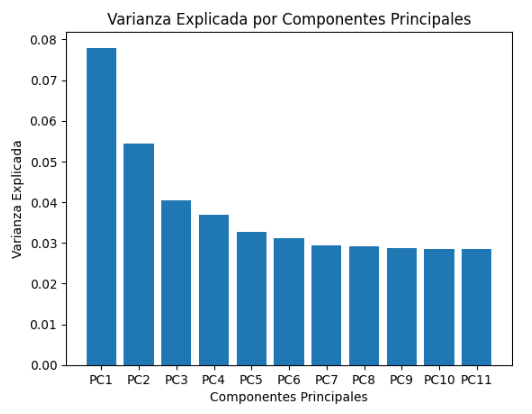


Imagen 4 Varianza Explicada por componentes principales

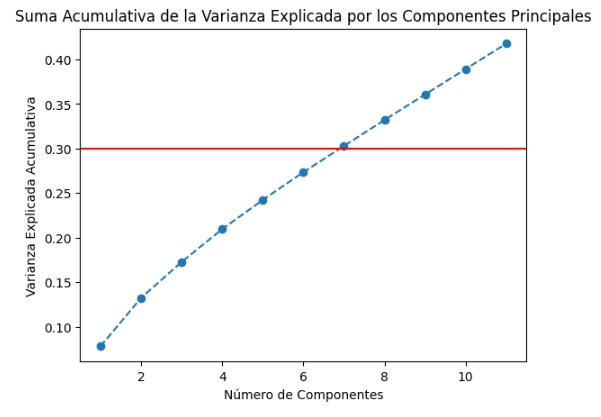


Imagen 5 Acumulacion de la varianza explicada por componentes principales

En la imagen 4 anterior podemos ver que con 4 componentes capturamos el 30% de la explicacion de la varianza por lo que corremos nuevamente el modelo, solo que esta vez seleccionamos 4 componentes y este seria nuestro modelo de PCA final.

n\_components=4

pca = PCA(n\_components=n\_components)

principal\_components = pca.fit\_transform(scaled\_data)

components=[]

n=1

while n<=n\_components:

    components.append("PC"+str(n))

    n=n+1

principal\_df = pd.DataFrame(data=principal\_components, columns=components)

explained\_variance = pca.explained\_variance\_ratio\_

print("Varianza Explicada:", explained\_variance)

Obtenemos como resultado lo siguiente:



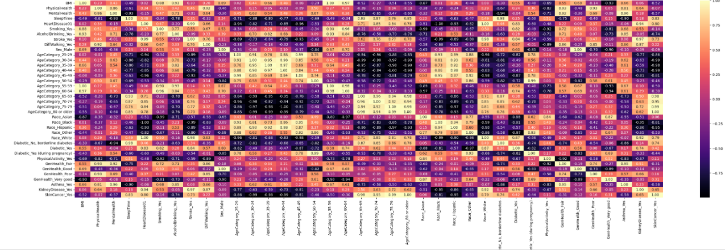
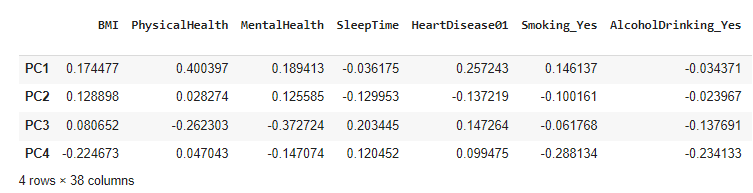


Imagen 6 Medida de Kendall’s tau

**3. Resultados y Discusión**

Interpretamos los resultados obtenidos, se analizará las contribuciones de cada variable original en cada componente principal. Los valores mas altos en magnitud indican una mayor contribución de esa variable a la formación del componente principal.



# 4. Conclusiones

El PCA, o Análisis de Componentes Principales, es una herramienta ampliamente utilizada y poderosa en ciencia de datos y análisis multivariante. Aunque tiene sus propias restricciones, el PCA ha demostrado ser valioso en diversas aplicaciones, desde la visualización y exploración de datos hasta la mejora del rendimiento de modelos de aprendizaje automático. Sin embargo, como ocurre con cualquier técnica estadística, es importante considerar cuidadosamente las suposiciones y limitaciones del PCA antes de aplicarlo a un conjunto de datos específico.

# Referencias

Joaquín Amat Rodrigo. (s.f.). Análisis de Componentes Principales (Principal Component Analysis, PCA) y t-SNE. Ciencia de datos, teoría y ejemplos prácticos en R y Python.

<https://www.cienciadedatos.net/documentos/35_principal_component_analysis>

¿Qué es el Análisis de Componentes Principales y cómo reducir el tamaño de una base de datos? (s.f.). Blog de Hiberus Tecnología. <https://www.hiberus.com/crecemos-contigo/analisis-de-componentes-principales/>

RPubs - Análisis de componentes principales (PCA). (s.f.). RPubs.

<https://rpubs.com/Cristina_Gil/PCA>

¿Qué es PCA en estadística? | KeepCoding Bootcamps. (s. f.). Recuperado 19 de junio de 2023, de <https://keepcoding.io/blog/que-es-pca-en-estadistica/>