FUTURA
ANALITICA AVANZADA

Webinar:

Análisis de Componentes Principales

Programa: Data Science for Business

Inicio : 12 agosto
Hora : 7:30-10:30pm (GMT-5 Lima)
Modo Online
Contacto : +51 908 814 045
Email : informes@futuradata.pe

FUTURA
ANALITICA AVANZADA

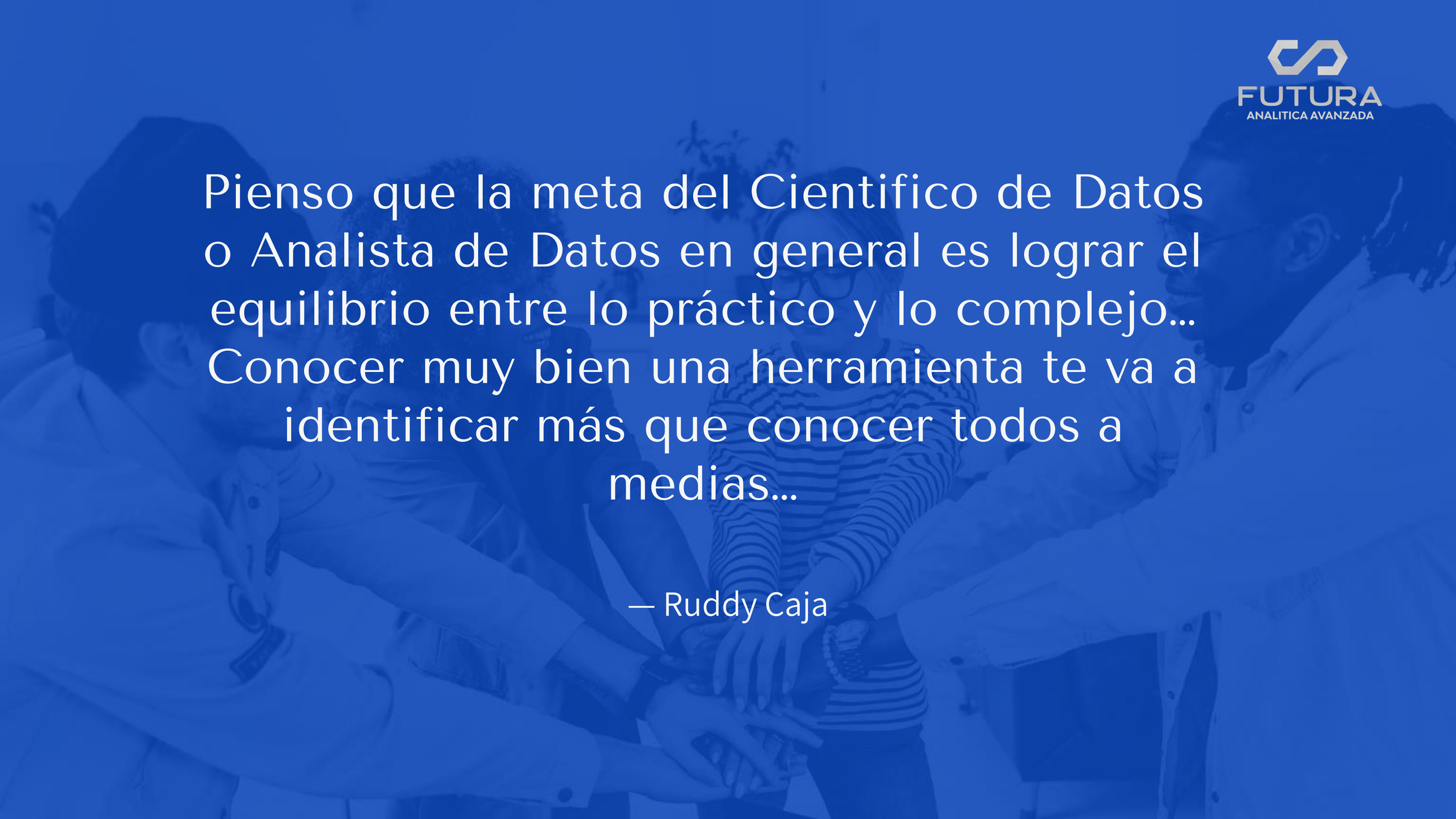
Hola!

UPC
Universidad Peruana
de Ciencias Aplicadas

CEO Futura

Ing. Ruddy Caja,

- ✓ Ing. Estadístico e Informático - UNALM
- ✓ Diploma Business Intelligence Specialist – UPC
- ✓ Autor: Cálculo de Probabilidades: Un enfoque teórico y práctico
- ✓ Experto en Analítica Inmobiliaria
- ✓ CEO Futura
- ✓ 12 años de experiencia en Banca (Inteligencia Comercial, Riesgos y Banca de Negocios)



Pienso que la meta del Cientifico de Datos o Analista de Datos en general es lograr el equilibrio entre lo práctico y lo complejo... Conocer muy bien una herramienta te va a identificar más que conocer todos a medias...

— Ruddy Caja

FUTURA
ANALITICA AVANZADA

Nuestro team de Mentores

Luis Felipe Garayar

Mentor Máster

Luis Angel Torres

Mentor Senior

Miguel A. Echeverre

Mentor Senior

Jonattan Ramos

Mentor Senior

Temas

- ¿De qué trata el PCA?
- Fundamento matemático
- Caso de uso
- Nuestro programa

FUTURA
ANALITICA AVANZADA

¿De qué trata el PCA?

review

Cómo funciona el Análisis de Componentes Principales

Es una técnica estadística de síntesis de información, o reducción de dimensión (en base a COMPONENTES). El ACP se basa en combinaciones lineales y transforma las variables originales en otras que se llaman componentes, las cuales son NO CORRELACIONADAS tomadas dos a dos.

- El ACP permite pasar a un nuevo conjunto de variables – las componentes principales – que gozan de la ventaja de **estar incorrelacionadas entre sí**, y que, además pueden ordenarse de acuerdo a la información que contienen. Esto último es muy útil ya que en los casos de regresiones las variables no deben estar correlacionadas y si se usan el 100% de componentes, evitaríamos la multicolinealidad.
- En otro caso no se requiere evitar la multicolinealidad porque no es el fin o la estrategia realizar una regresión lineal, sino más bien es simplemente trabajar con una cantidad menor de variables que representen el mismo problema **minimizando la pérdida de información.**

La matemática detrás...

Se han observado p variables $X_1, X_2, X_3 \ldots X_p$ sobre una muestra de n individuos u observaciones. La matriz de datos muestrales es:

$$\mathbf{X} = \begin{bmatrix} x_{11} x_{12} \cdots x_{1p} \\ x_{21} x_{22} \cdots x_{2p} \\ \vdots \quad \vdots \quad \ddots \quad \vdots \\ x_{n1} x_{n2} \cdots x_{np} \end{bmatrix}$$

En adelante supondremos que $\mathbf{X}$ es una "matriz centrada" (que cada observación de la variable ha sido sustraída de su media muestral y que la media de la variable conjunta es cero)

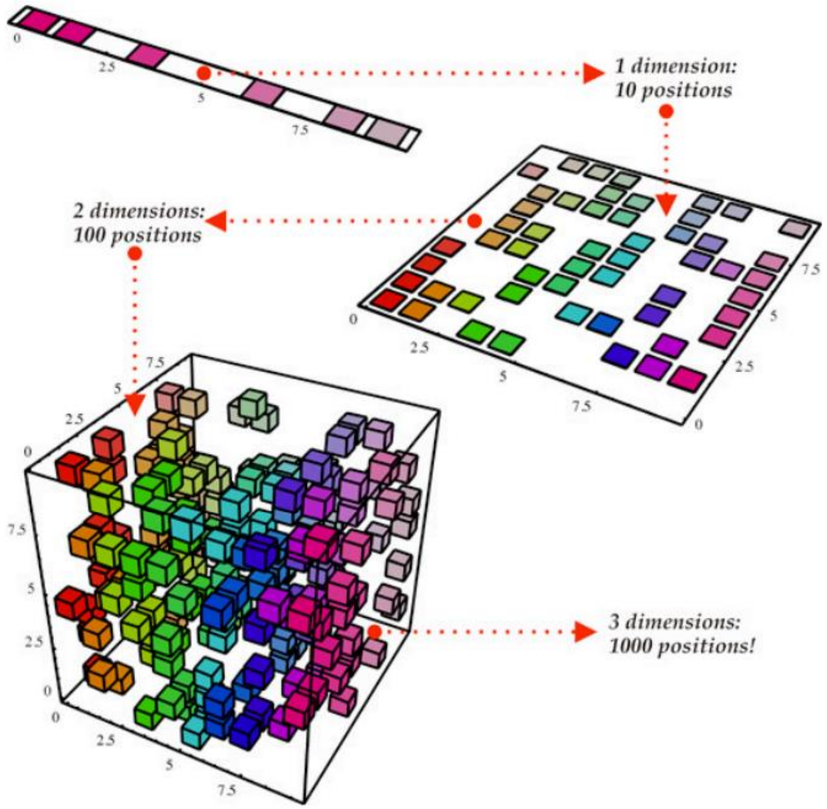
Problema: ¿Podemos describir la "información" contenida en estos datos mediante algún conjunto de variables menor que el de variables originales?
Idea: Si una variable es función de otras, contiene información redundante.

La matemática detrás...

Por tanto, si las p variables observadas están fuertemente correlacionadas, será posible sustituirlas por menos variables sin gran pérdida de “información”.

Esta reducción de la dimensión va a permitir:

- Simplificar posteriores análisis, que se harán a partir de un menor número de variables que el original.
- Una representación gráfica de los individuos en dimensión reducida (generalmente, 1 ó 2).
- Examinar e interpretar las relaciones entre las variables observadas.



La matemática detrás...

RECUERDA: TEOREMA DE LA DESCOMPOSICIÓN ESPECTRAL

Toda matriz simétrica $\mathbf{A}$ de orden $\boldsymbol{pxp}$ puede ser escrita como un producto de matrices:

$$\mathbf{A} = \mathbf{\Gamma \Lambda \Gamma'}$$

$\mathbf{\Gamma}$: Matriz ortogonal pxp contiene los **autovectores o vectores propios** de $\mathbf{A}$

$\mathbf{\Lambda}$**:** Matriz diagonal pxp contiene los **autovalores o valores propios** de $\mathbf{A}$

Definición y obtención de las Componentes Principales

Sean $\mathbf{X} = [X_1, \ldots, X_p]$ y $\mathbf{S} = \text{var}(\mathbf{X})$ su matriz de covarianzas.

Puesto que $\mathbf{S} \geq \mathbf{0}$ y simétrica, su descomposición espectral es:

$$\mathbf{S} = \mathbf{T}\Lambda\mathbf{T}'$$

Donde $\mathbf{T}'\mathbf{T} = \mathbf{T}\mathbf{T}' = \mathbf{I}$, con $\mathbf{T} = [\mathbf{t}_1, \ldots, \mathbf{t}_p]$ y $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$,

Las componentes principales de $\mathbf{X}$ son las nuevas variables

$$\mathbf{Y}_j = \mathbf{X}\mathbf{t}_j, \qquad j = 1, 2, \ldots, p$$

Para cada j, la nueva variable $\mathbf{Y}_j$ se construye a partir del *j-ésimo* autovalor de $\mathbf{S}$.

Propiedades de las Componentes Principales

- Las componentes principales tienen **varianza decreciente**.

$$\left.\begin{aligned} \text{var}(Y_1) = \text{var}(\mathbf{Xt}_1) = \mathbf{t}_1'\mathbf{St}_1 &= \lambda_1\mathbf{t}_1'\mathbf{t}_1 = \lambda_1 \\ \text{var}(Y_2) = \text{var}(\mathbf{Xt}_2) = \mathbf{t}_2'\mathbf{St}_2 &= \lambda_2\mathbf{t}_2'\mathbf{t}_2 = \lambda_2 \\ &\vdots \\ \text{var}(Y_p) = \text{var}(\mathbf{Xt}_p) = \mathbf{t}_p'\mathbf{St}_p &= \lambda_p\mathbf{t}_p'\mathbf{t}_p = \lambda_p \end{aligned}\right\} \text{Con } \lambda_1 > \lambda_2 > \cdots > \lambda_p$$

- y, están **incorrelacionadas** unas con otras.

$$\text{cov}(Y_i, Y_j) = \text{cov}(\mathbf{Xt}_i, \mathbf{Xt}_j) = \mathbf{t}_i'\mathbf{St}_j = \lambda_j\mathbf{t}_i'\mathbf{t}_j = 0, \text{para } i \neq j,$$

puesto que $\mathbf{T}$ es una matriz ortogonal

Propiedades de las Componentes Principales

- Las covarianzas entre cada componente principal y las variables originales $\mathbf{X}_i$ son:

  $\text{cov}(Y_j, [X_1, \ldots, X_p]) = \lambda_j \mathbf{t}'_j = 0, \text{para } j = 1, 2, ..., p$

  Utilizando que $\mathbf{Y} = \mathbf{XT}$ y la descomposición espectral de $\mathbf{S}$

  $\text{cov}(\mathbf{Y},\mathbf{X}) = \frac{1}{n}\mathbf{Y}'\mathbf{X} = \frac{1}{n}\mathbf{T}'\mathbf{X}'\mathbf{X} = \mathbf{T}'\mathbf{S} = \mathbf{T}'(\mathbf{T}\Lambda\mathbf{T}') = \Lambda\mathbf{T}'$

  La fila j de la matriz proporciona las covarianzas entre Y_j y las variables originales $X_1, X_2, \ldots, X_p$.

  Por ejemplo, las covarianzas entre Y_1 y $X_1, \ldots, X_p$ es $\lambda_1 \mathbf{t}'_1$

- La correlación entre Y_j y la variable original X_i es:

$$corr(Y_j, X_i) = \frac{cov(Y_j, X_i)}{\sqrt{var(Y_j)var(X_i)}} = \frac{\lambda_j t_{ij}}{\sqrt{\lambda_j s_{ii}}} = t_{ij}\sqrt{\frac{\lambda_j}{s_{ii}}}$$

  Donde t_{ij} es el elemento *i-ésimo* del autovector $\mathbf{t}_j$

Representación de las observaciones

Con las nuevas coordenadas dadas por los componentes principales, el individuo i-ésimo, es decir, la fila $\mathbf{x}_i^{'}=(x_{i1},\ldots,x_{ip})$ de la matriz de datos $\mathbf{X}$, se expresa como:

$$\mathbf{y}_i' = \mathbf{x}_i'\mathbf{T} = (\mathbf{x}_i'\mathbf{t}_1, \ldots, \mathbf{x}_i'\mathbf{t}_p)$$

La matriz de datos con transformaciones es $\mathbf{Y} = \mathbf{XT}$, que representa las "observaciones" de las nuevas variables (componentes principales) sobre los ***n*** individuos de la muestra.

Esta transformación puede interpretarse geométricamente **considerando los *n* individuos como *n puntos*** en el **espacio** $\mathbb{R}^p$.

Consideremos la distancia euclídea (al cuadrado) entre los individuos *i-ésimo* y *j-ésimo*, en las nuevas coordenadas.

$$d^2_{Euclid}(i,j) = (\mathbf{y}_i' - \mathbf{y}_j')(\mathbf{y}_i - \mathbf{y}_j) = (\mathbf{x}_i'\mathbf{T} - \mathbf{x}_j'\mathbf{T})(\mathbf{T}'\mathbf{x}_i - \mathbf{T}'\mathbf{x}_j)$$
$$d^2_{Euclid}(i,j) = (\mathbf{x}_i' - \mathbf{x}_j')\mathbf{T}'\mathbf{T}(\mathbf{x}_i - \mathbf{x}_j) = (\mathbf{x}_i' - \mathbf{x}_j')(\mathbf{x}_i - \mathbf{x}_j)$$

Ignorando orientaciones, podemos pensar la transformación como ***una rotación*** en $\mathbb{R}^p$

El primero de los nuevos ejes (la primera componente principal) es la dirección a lo largo de la cual la dispersión de los puntos-individuos es máxima. Sucesivamente, cada componente principal es aquella dirección, ortogonal a las anteriores, a lo largo de la cual hay dispersión máxima.

Reducción de la dimensión

La variación total de $\mathbf{X}$ se define como $tr(\mathbf{S}) = \sum_{I=1}^{p} \lambda_i$

La variación total de $\mathbf{Y} = \mathbf{XT}$ es igual a la variación total de $\mathbf{X}$:

$$tr\big(var(\mathbf{Y})\big) = tr\left(\frac{1}{n}\mathbf{T}'\mathbf{X}'\mathbf{X}\mathbf{T}\right) = tr(\mathbf{T}'\mathbf{S}\mathbf{T}) = tr(\mathbf{T}'\mathbf{T}\boldsymbol{\Lambda}\mathbf{T}'\mathbf{T}) = \sum_{i=1}^{p} \lambda_i$$

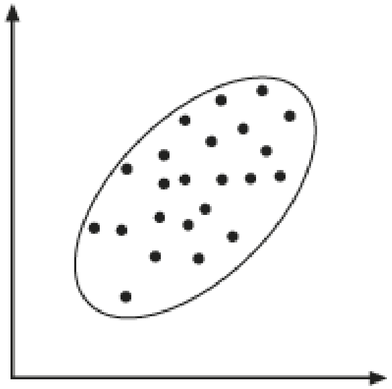
Puesto que, $\mathbf{S} = \mathbf{T}\boldsymbol{\Lambda}\mathbf{T}'$, $\mathbf{T}$ es una matriz ortogonal.

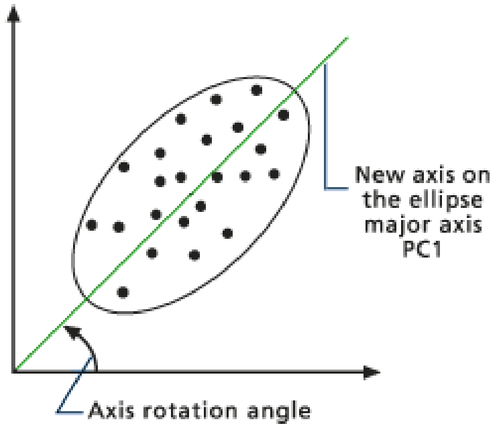
Cuando el cociente (**porcentaje de variabilidad explicada**)

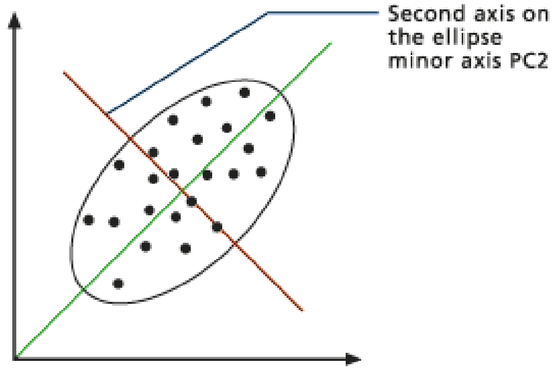
$$P_q = \frac{\sum_{i=1}^{p} \lambda_i}{tr\mathbf{S}} \text{x} 100, \qquad q < p,$$

es cercano a 100%, entonces las variables $Y_1, Y_2, \ldots, Y_q$ pueden reemplazar a $X_1, X_2, \ldots, X_p$ sin gran pérdida de información, en términos de “variación total”.

Reducción de la dimensión







Sean $\mathbf{X} = [X_1, \ldots, X_p]$ y $\mathbf{S} = \text{var}(\mathbf{X})$ su matriz de covarianzas.

$$\mathbf{S} = \mathbf{T}\Lambda\mathbf{T}'$$

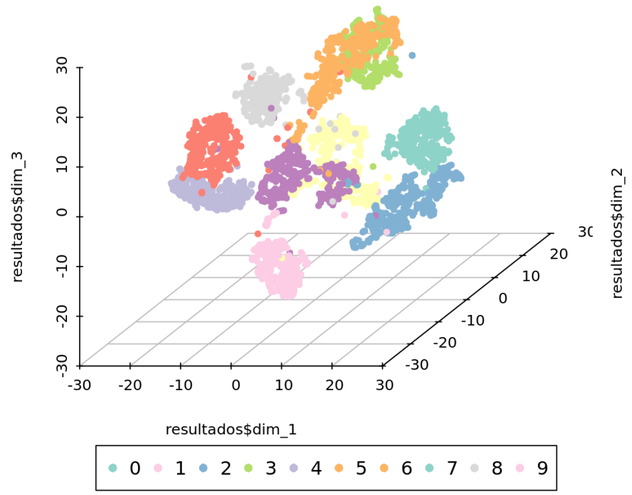
$$\mathbf{Y}_j = \mathbf{X}\mathbf{t}_j, \quad j = 1, 2, \ldots, p$$

$$\begin{aligned}
\text{var}(Y_1) = \text{var}(\mathbf{X}\mathbf{t}_1) &= \mathbf{t}_1'\mathbf{S}\mathbf{t}_1 = \lambda_1\mathbf{t}_1'\mathbf{t}_1 = \lambda_1 \\
\text{var}(Y_2) = \text{var}(\mathbf{X}\mathbf{t}_2) &= \mathbf{t}_2'\mathbf{S}\mathbf{t}_2 = \lambda_2\mathbf{t}_2'\mathbf{t}_2 = \lambda_2 \\
&\vdots \\
\text{var}(Y_p) = \text{var}(\mathbf{X}\mathbf{t}_p) &= \mathbf{t}_p'\mathbf{S}\mathbf{t}_p = \lambda_p\mathbf{t}_p'\mathbf{t}_p = \lambda_p
\end{aligned}$$

Con $\lambda_1 > \lambda_2 > \cdots > \lambda_p$

Otros métodos para reducir dimensionalidad

- PCA (linear)
- t-SNE (non-parametric/ nonlinear)
- Sammon mapping (nonlinear)
- Isomap (nonlinear)
- LLE (nonlinear)
- CCA (nonlinear)
- SNE (nonlinear)
- MVU (nonlinear)
- Laplacian Eigenmaps (nonlinear)



Caso de Uso

Analicemos...

FUTURA
ANALITICA AVANZADA

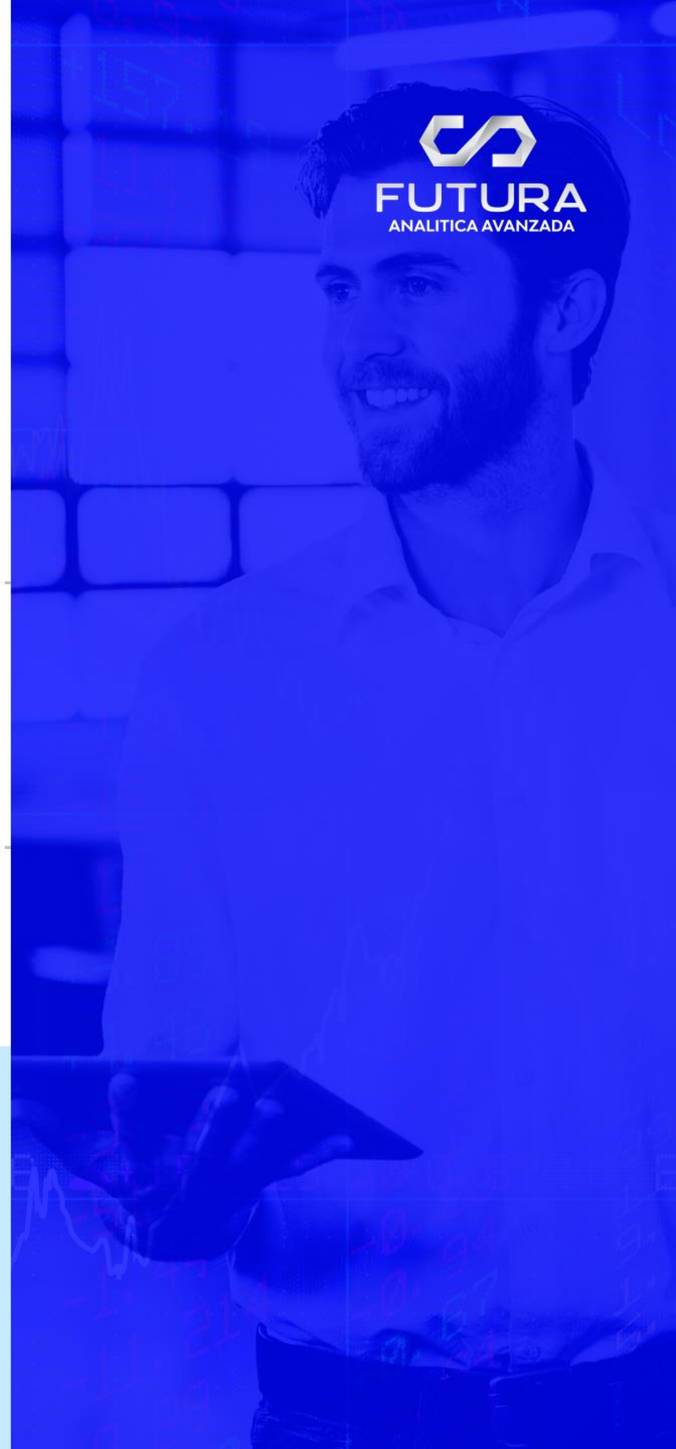
Nuestro programa académico

ROADMAP DE APRENDIZAJE

72 HORAS
(3h por sesión, dos veces a la semana)

- INTRODUCCIÓN A LA CIENCIA DE DATOS
- PYTHON PARA CIENCIA DE DATOS
- VISUALIZACIÓN Y ANÁLISIS DE DATOS
- TRANSFORMACIÓN DE DATOS Y WOEs
- ANÁLISIS EXPLORATORIO DE DATOS
- MODELOS DE REGRESIÓN
- MODELO LOGIT & KNN
- TECNICAS DE CLASIFICACIÓN (RANDOM FOREST)
- MODELOS DE SEGMENTACIÓN (K-MEANS Y CLUSTER JERÁRQUICO)
- COMPONENTES PRINCIPALES Y REDUCCIÓN DE DIMENSIONES
- SUPPORT VECTOR MACHINE & MODELOS ENSAMBLADOS
- MARKET BASKET ANALYSIS
- TEXT MINING & DEEP LEARNING
- ANÁLISIS DE SERIES DE TIEMPO
- COMPARACIÓN DE MODELOS

FUTURA
ANALITICA AVANZADA



CONTENIDO

MÓDULO 1
CIENCIA DE DATOS & HERRAMIENTAS

12h

- Introducción a la ciencia de datos
- Proceso de un proyecto de ciencia de datos
- Programación en Python

MÓDULO 2
PRINCIPIOS BÁSICOS DE LA CIENCIA DE DATOS

18h

- Visualización de datos
- Análisis exploratorio de datos
- Tratamiento e imputación de datos perdidos
- Limpieza de datos: Outliers univariados y multivariados
- Pre-procesamiento de datos
- Transformación de datos y WOEs (Weight of Evidence)
- Selección inicial de variables

MÓDULO 3
ALGORITMOS DE LA CIENCIA DE DATOS

21h

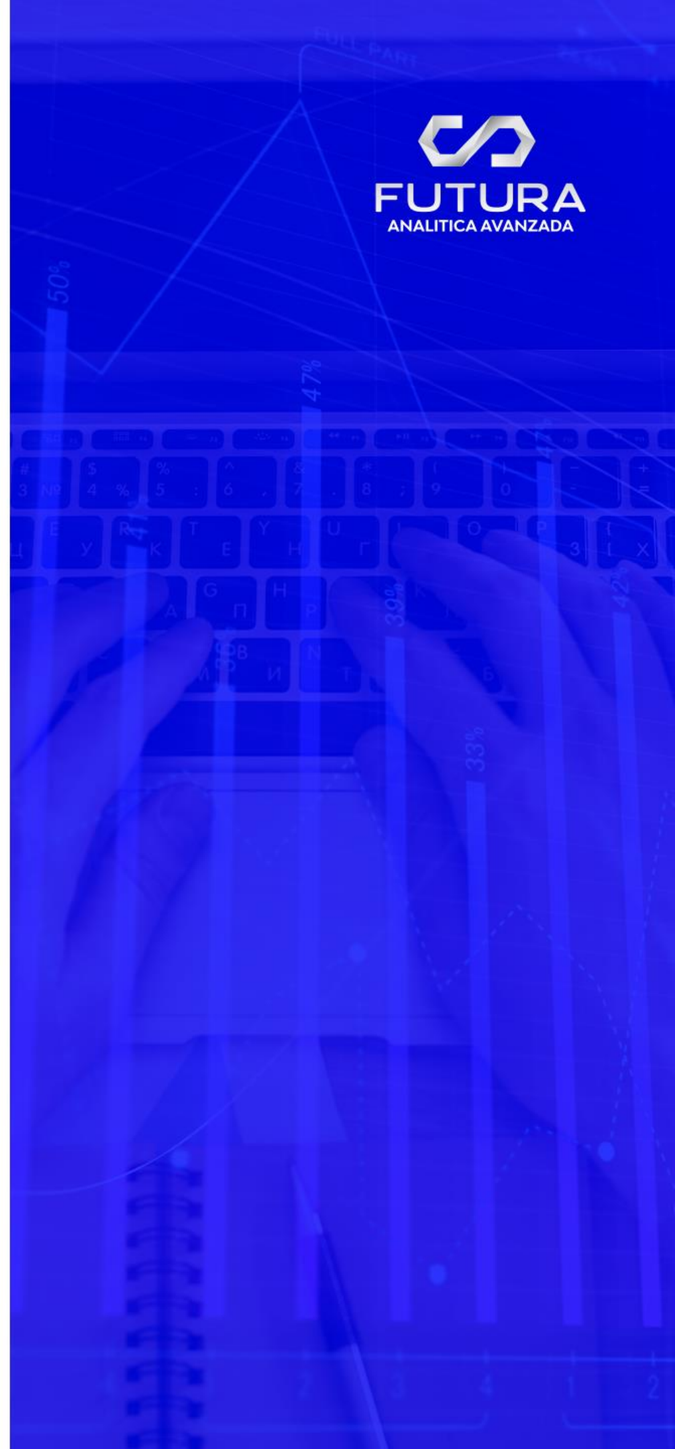
- Regresión Lineal y técnicas de regularización
- Regresión Logística y K-Nearest Neighbors (KNN)
- Árboles de Regresión y Clasificación
- Evaluación y comparación de modelos
- K-Means y técnicas de clustering jerárquico
- Componentes Principales y reducción de dimensiones

MÓDULO 4
TÉCNICAS AVANZADAS

21h

- Support Vector Machine (SVM)
- Modelos ensamblados: Bagging, Random Forest y Boosting
- Market Basket Analysis, sistemas de recomendación
- Introducción al Text Mining
- Deep Learning: Redes Neuronales, Convolucionales y su aplicación
- Análisis de series de tiempo

FUTURA
ANALITICA AVANZADA



CASOS DE ESTUDIO

FINANCIERO, SEGUROS Y COBRANZAS

- Marketing Bancario
- Predictor de Ingresos
- Credit Scoring
- Fuga de Clientes Telco
- Seguro: Pago de Primas de renovación
- Contactabilidad Cobranzas

SALUD Y RECURSOS HUMANOS

- Health Analytics
- Human Resources Analytics

RETAIL

- Big Mart Analytics
- Black Friday

INMOBILIARIO

- Análisis de Precios de Viviendas
- Adquisición de Créditos Hipotecarios

Rueda de preguntas...

FUTURA
ANALITICA AVANZADA

Programa: Data Science for Business

Inicio : 12 agosto
Hora : 7:30-10:30pm (GMT-5 Lima)
Modo Online
Contacto : +51 908 814 045
Email : informes@futuradata.pe