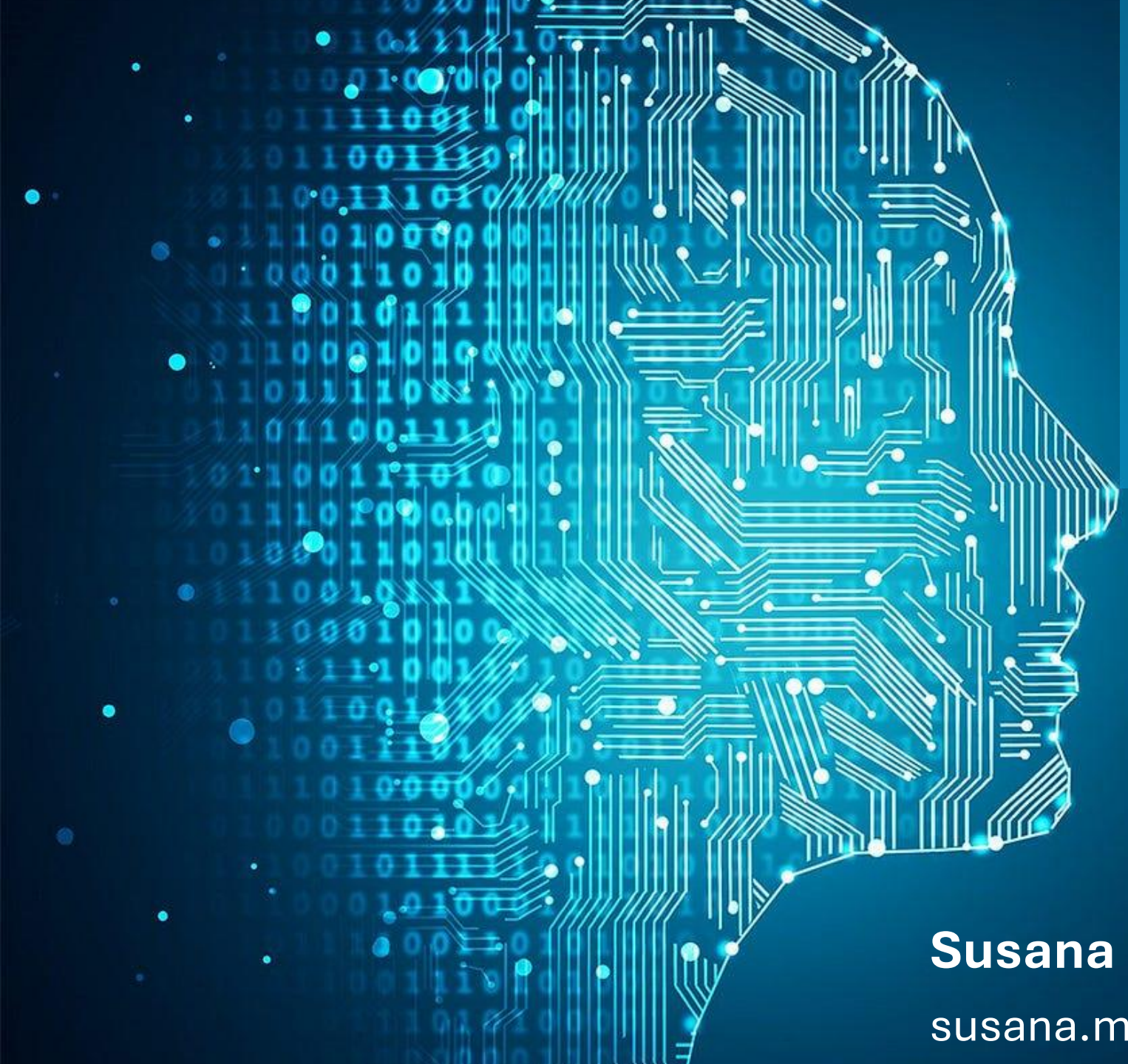
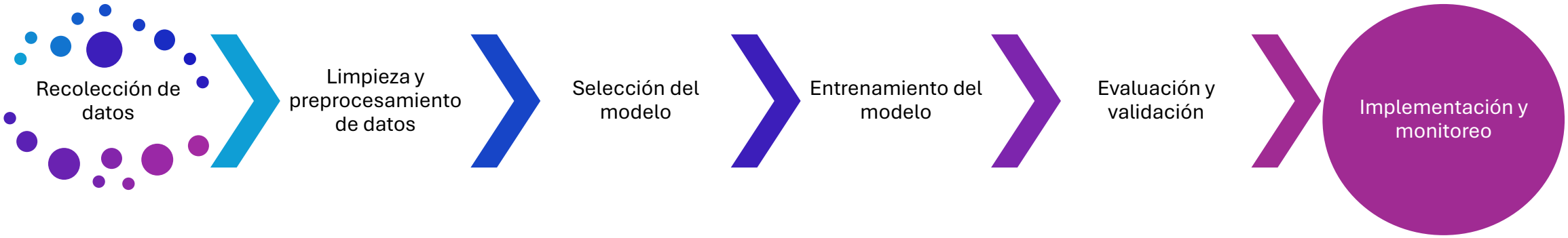


Machine Learning

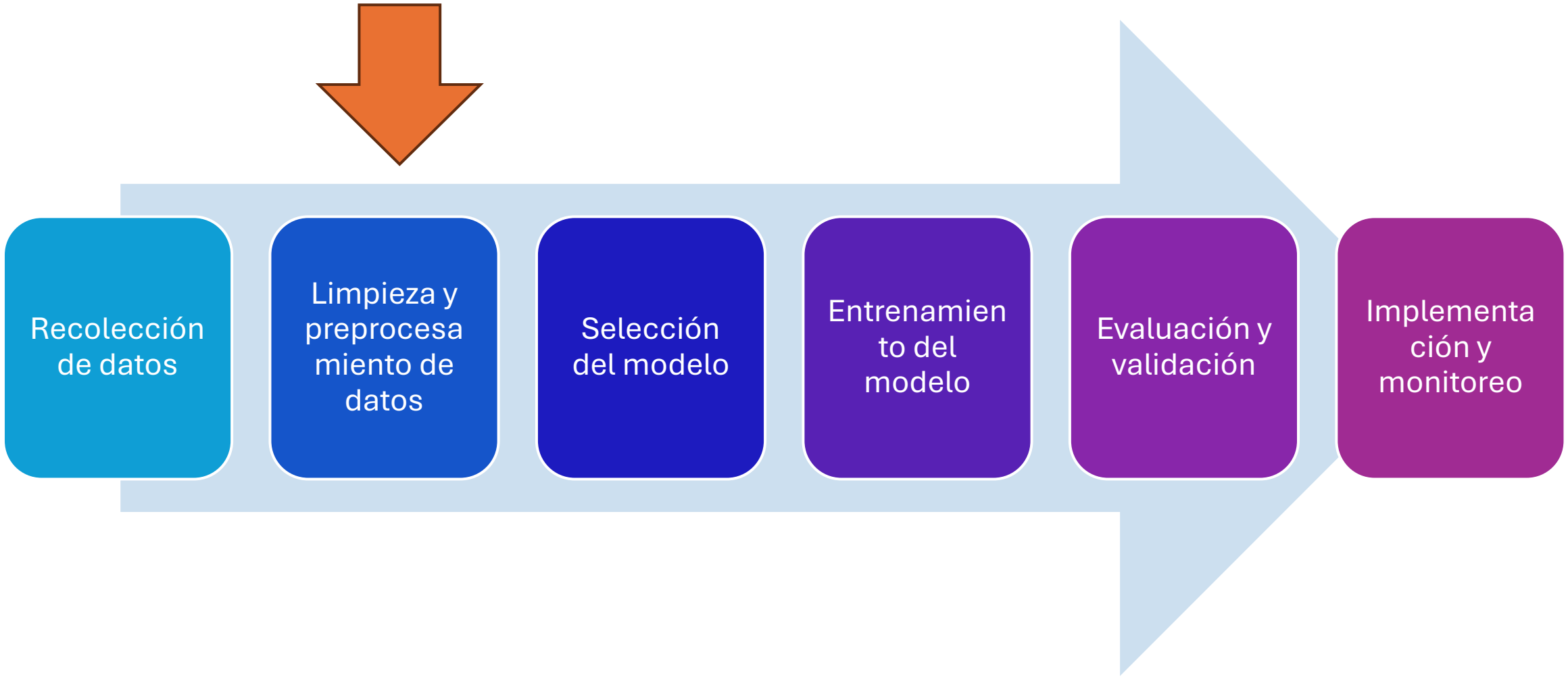
Susana Medina Gordillo

susana.medina@correounivalle.edu.co

Flujo de trabajo en Machine Learning



Flujo de trabajo en Machine Learning



Selección y Transformación de Características

Selección de Características: Introducción

- Selección y transformación de características
- Reducción de la dimensionalidad

Selección de Características: Introducción

Objetivos de la sesión

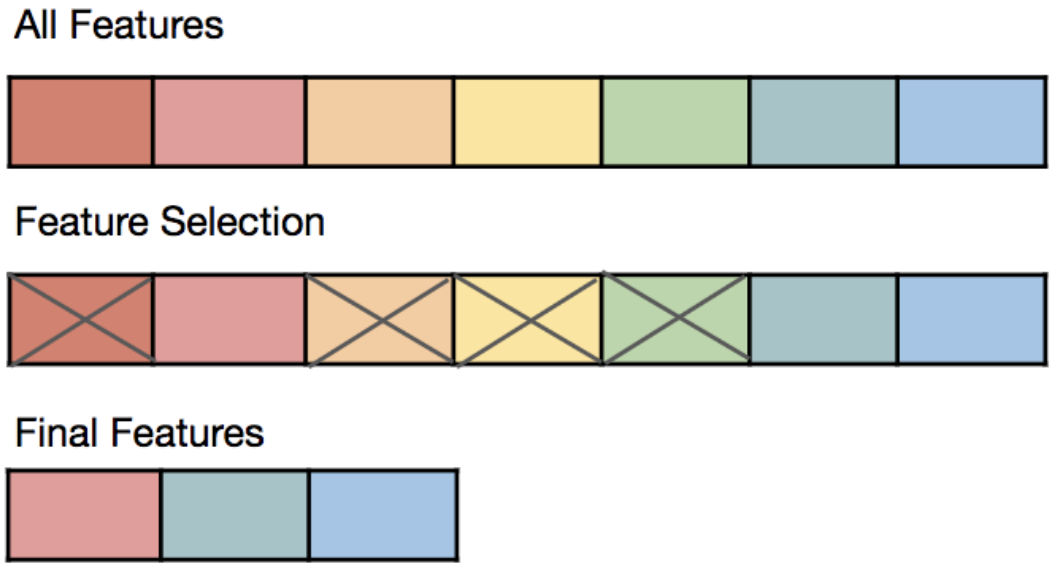
- Comprender las técnicas de selección de características.
- Aprender a codificar variables categóricas.
- Conocer las técnicas de reducción de dimensionalidad

¿Por qué es importante la selección y transformación de características?

- **Rendimiento del modelo:** Reducir la dimensionalidad y eliminar ruido mejora la precisión.
- **Interpretabilidad:** Modelos más simples son más fáciles de entender.
- **Eficiencia computacional:** Menos características aceleran el entrenamiento y la predicción.

Selección de Características

¿Qué es la Selección de Características?



Feature Selection

Es el **proceso** de elegir el subconjunto más relevante de características de un conjunto de datos.

Métodos de Selección de Características

Métodos de filtro	○ Basados en estadísticas (p. ej., chi-cuadrado, ANOVA).
Métodos de envoltura	○ Basados en el rendimiento del modelo (p. ej., selección recursiva de características).
Métodos incrustados	○ Integrados en el algoritmo de aprendizaje (p. ej., regularización L1).

Beneficios de la selección de caracteristicas

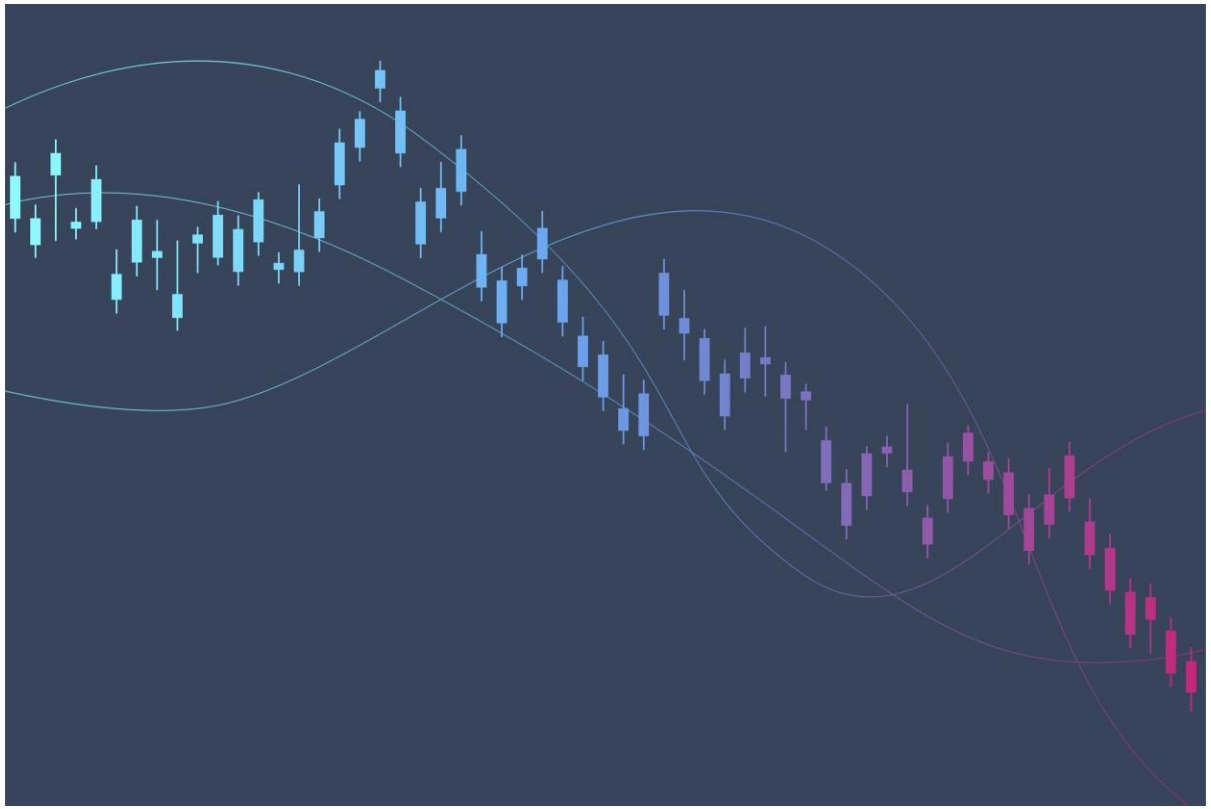
Evitar el sobre ajuste

Mejorar la precisión

Reducir el tiempo de entrenamiento

Codificación Categórica

Es el **proceso** de convertir variables categóricas en representaciones numéricas.



Técnicas de codificación

Codificación ordinal

- Asigna **números enteros** basados en el orden de las categorías.

Codificación de etiquetas

- Asigna un **número entero** a cada categoría.

Codificación one-hot

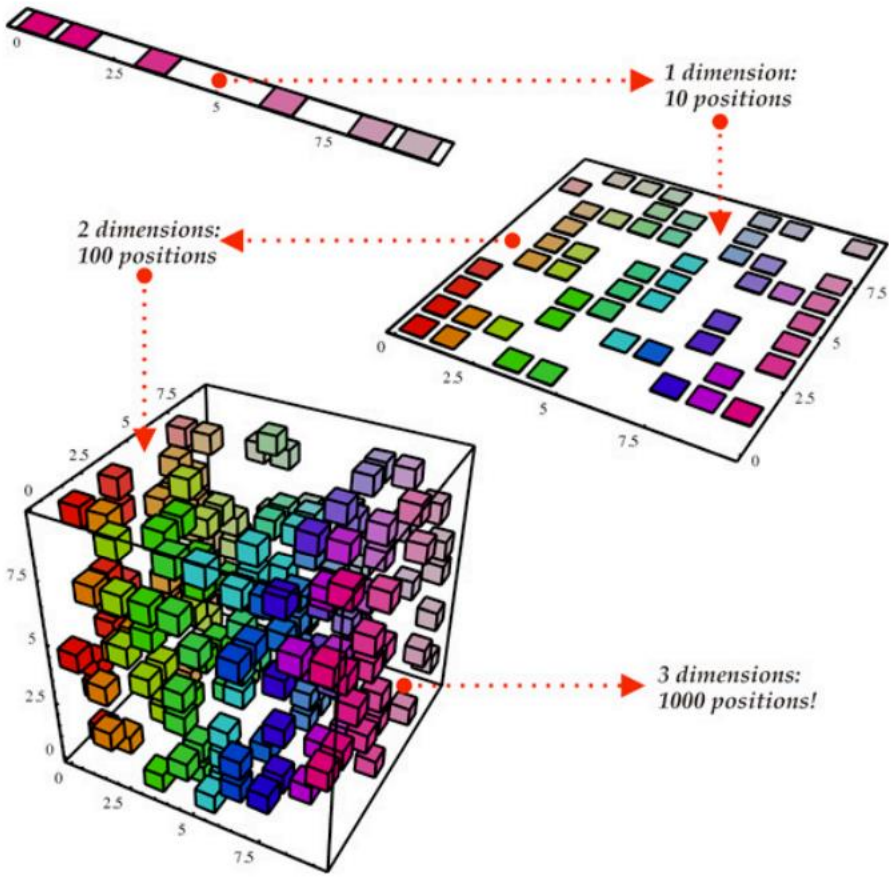
- Crea columnas **binarias** para cada categoría.

Generación de nuevas variables

- Combinación de características
- Características polinómicas

Reducción de Dimensionalidad

¿Qué es la Reducción de Dimensionalidad?



Dimensionality Reduction

Es el **proceso** de **reducir** el número de variables en un conjunto de datos

Técnicas de Reducción de Dimensionalidad

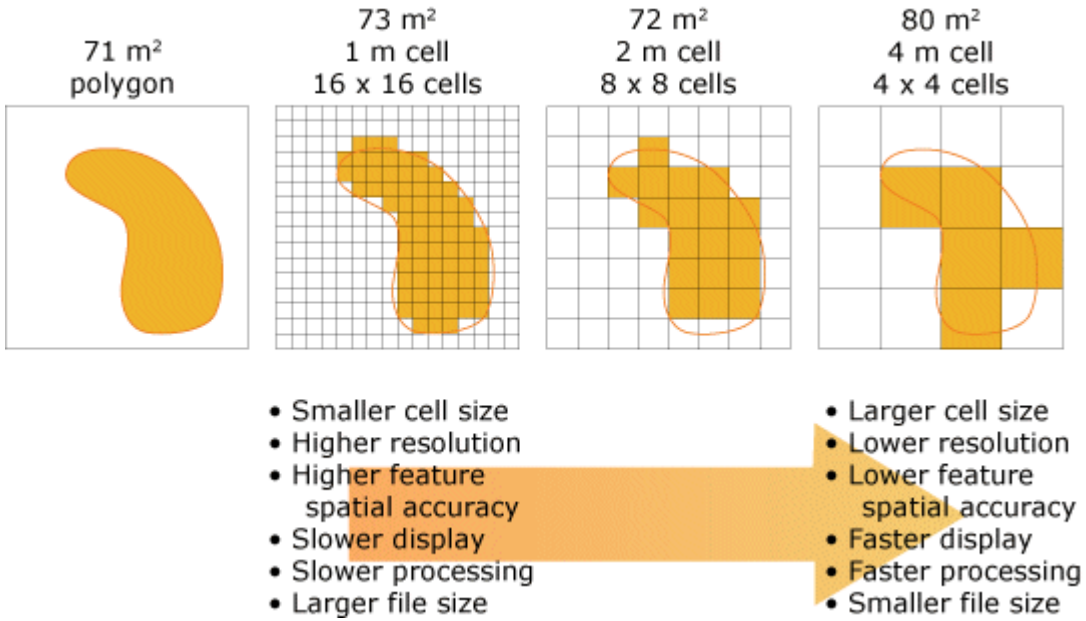
Análisis de componentes principales (**PCA**)	○ Transforma las características en componentes principales.
t-SNE (t-distributed Stochastic Neighbor Embedding)	○ Reduce la dimensionalidad para visualización.
UMAP (Uniform Manifold Approximation and Projection)	○ Similar a t-SNE, pero con algunas mejoras de rendimiento.

Cómo hacer selección de características en python?

- ✓ Selección de **características** con *SelectKBest*
- ✓ Codificación **one-hot** con *OneHotEncoder*
- ✓ Reducción de **dimensionalidad** con *PCA*

Selección de Características : Casos de uso

- ✓ **Análisis de texto**: Reducción de la dimensionalidad del vocabulario.
- ✓ **Genómica**: Selección de genes relevantes para una enfermedad.
- ✓ **Imágenes**: Reducción de las dimensiones de los pixeles.



scikit-learn

Machine Learning in Python

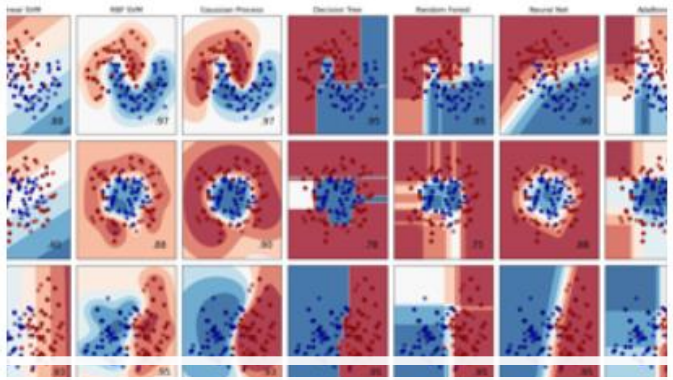
Getting Started | Release Highlights for 1.6

- Simple and efficient tools for predictive data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

Identifying which category an object belongs to.

Applications: Spam detection, image recognition.
Algorithms: Gradient boosting, nearest neighbors, random forest, logistic regression, and more...

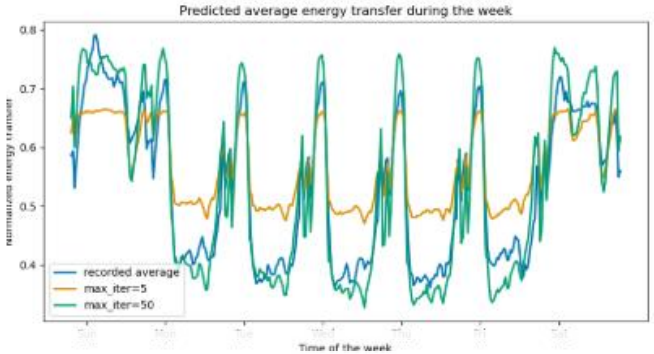


Examples

Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, stock prices.
Algorithms: Gradient boosting, nearest neighbors, random forest, ridge, and more...

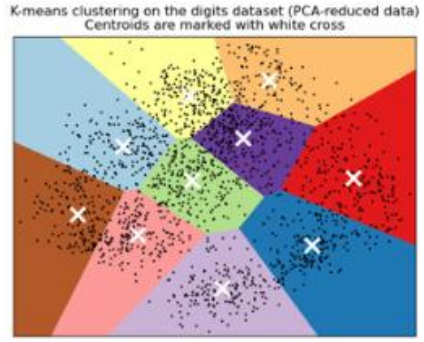


Examples

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, grouping experiment outcomes.
Algorithms: k-Means, HDBSCAN, hierarchical clustering, and more...



Examples

Dimensionality reduction

Reducing the number of random variables to consider.

Applications: Visualization, increased efficiency.

Model selection

Comparing, validating and choosing parameters and models.

Preprocessing

Feature extraction and normalization.

Applications: Transforming input data such as text for

Conclusiones...

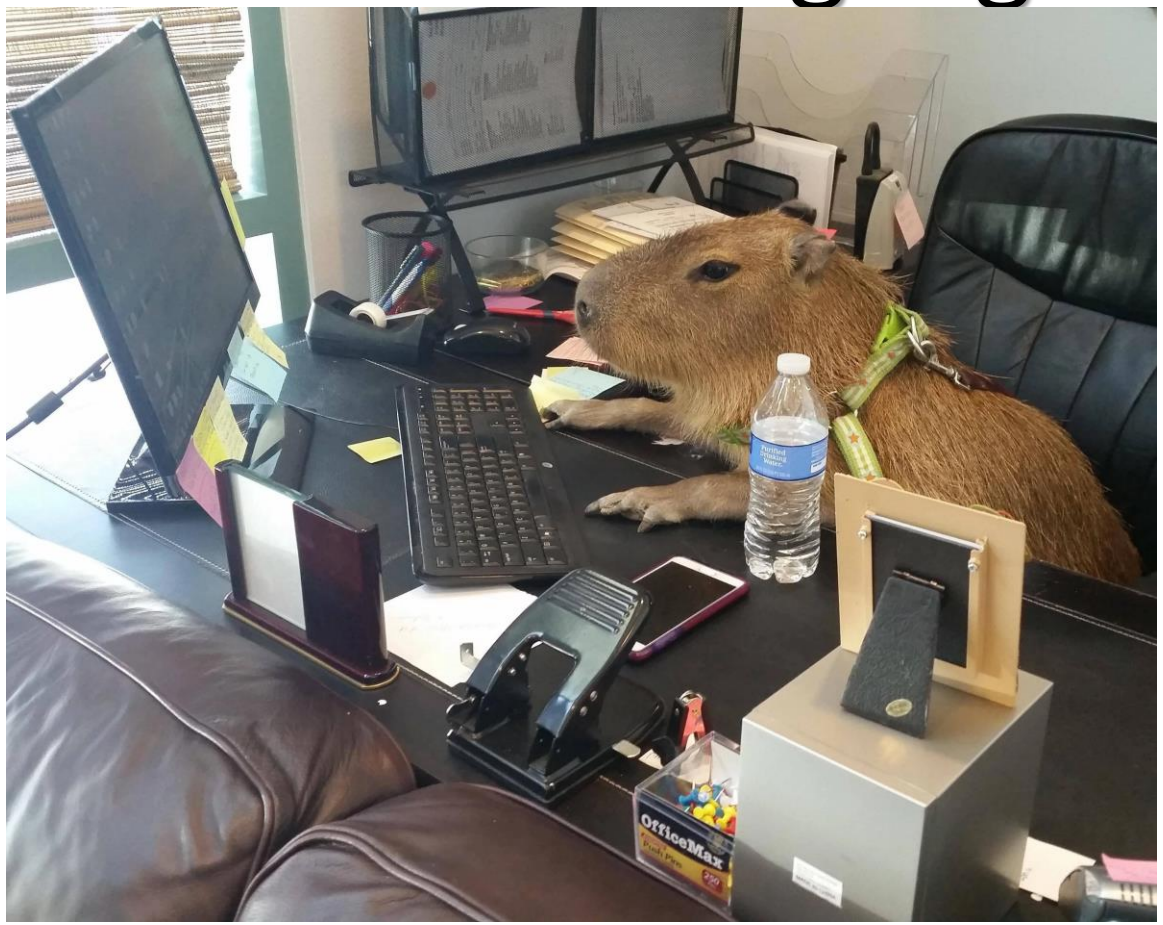
La selección y transformación de características son esenciales para el éxito del machine learning.

Existen diversas técnicas para abordar diferentes tipos de datos y problemas.

La práctica y la experimentación son fundamentales para dominar estas técnicas.

Ejercicio práctico

colab.research.google.com

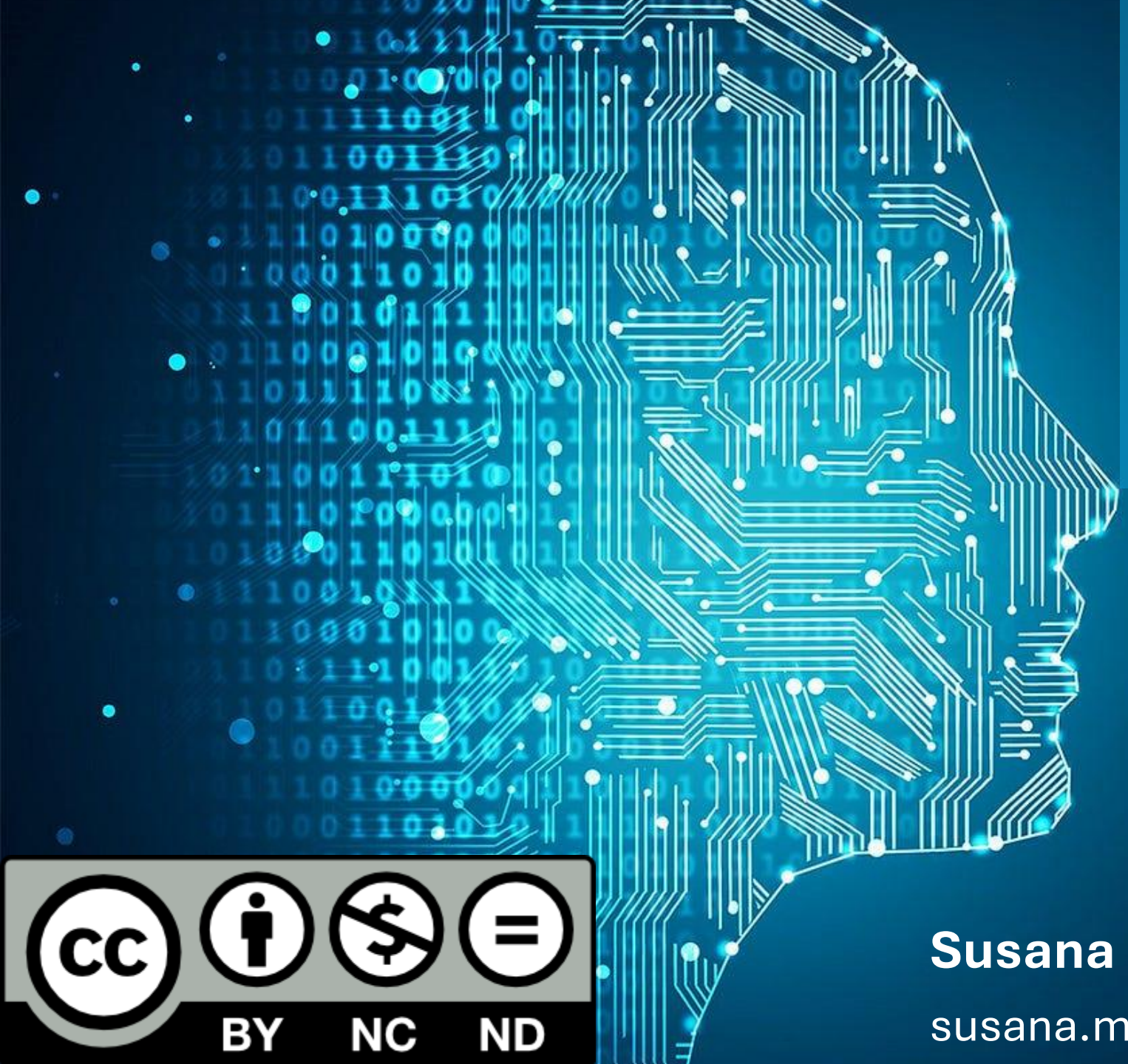


Ejercicio práctico: Google Colaboratory (*Colabs*)

- Página oficial: https://colab.google/
- Abrir Colab (incluye tutorial): https://colab.research.google.com/
- Guía para EDA: https://colab.research.google.com/github/Tanu-N-Prabhu/Python/blob/master/Exploratory_data_Analysis.ipynb
- Guía / tutorial para Selección de características con **scikit-learn**: https://www.datacamp.com/tutorial/feature-selection-python

Referencias

- “scikit-learn: Machine Learning in Python”. Consultado: el 20 de febrero de 2025. [En línea]. Disponible en: https://scikit-learn.org/stable/
- Imagen de Features. Consultado: el 20 de febrero de 2025. [En línea]. Disponible en: https://themanoftalent.medium.com/feature-selection-9b1609f1f6b0
- Imagen de Transforming variables. Consultado: el 20 de febrero de 2025. [En línea]. Disponible en: https://www.datasklr.com/ols-least-squares-regression/transforming-variables
- “A Probabilistic Algorithm to Reduce Dimensions: t — Distributed Stochastic Neighbor Embedding (t-SNE)”. Consultado: el 20 de febrero de 2025. [En línea]. Disponible en: https://pub.towardsai.net/a-probabilistic-algorithm-to-reduce-dimensions-t-distributed-stochastic-neighbor-embedding-23ff457fbc8a
- Información e ideas presentadas basadas en el conocimiento general de modelos de lenguaje de IA. Gemini 2.9 Flash. Consultado: el 20 de febrero de 2025. [En línea].



Machine Learning

Susana Medina Gordillo

susana.medina@correounivalle.edu.co