Modelos de clasificación

Prof. Dr. Jorge Zavaleta
zavaleta.jorge@gmail.com

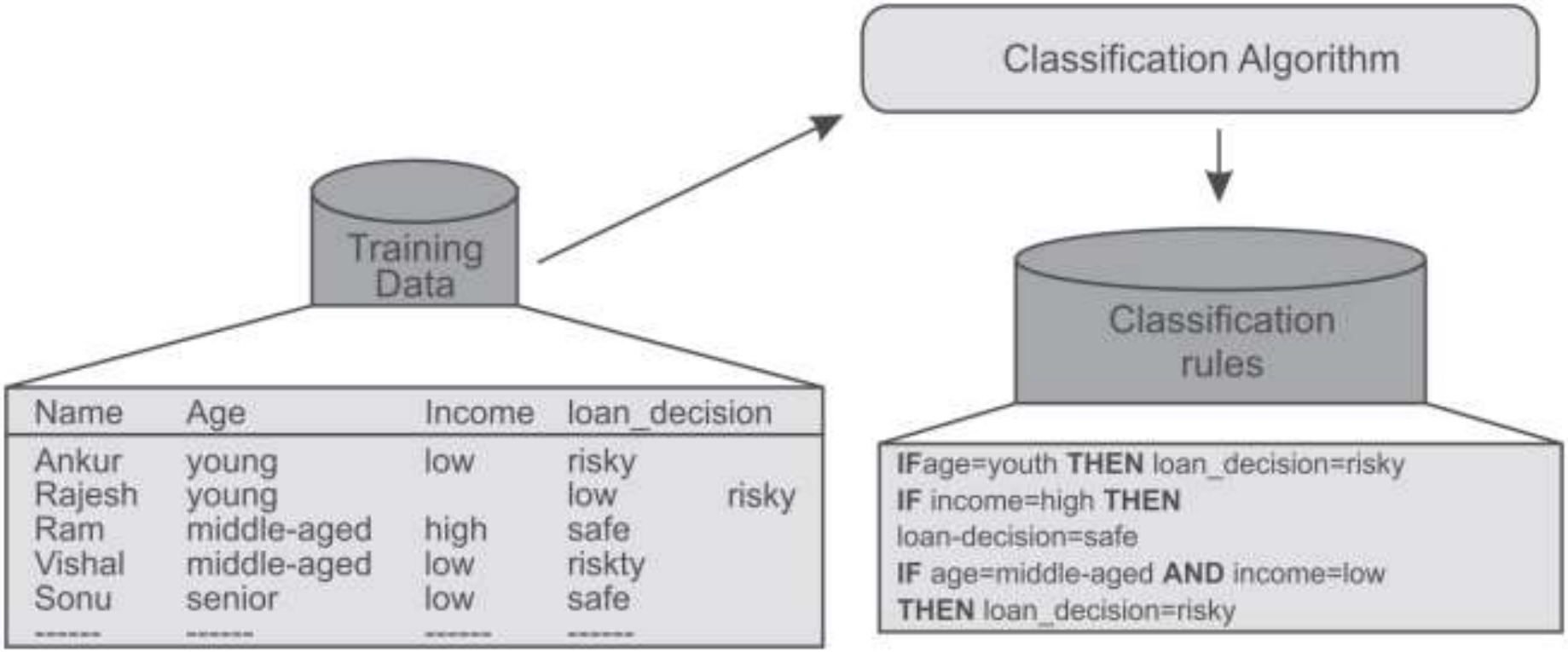
Modelos de clasificación

- En la minería de datos, los **modelos de clasificación** son herramientas estadísticas que aprenden a categorizar datos en diferentes clases o categorías.
- **La clasificación es el proceso de categorizar o agrupar datos en diferentes clases o categorías en función de determinadas características o atributos**.
- Esta técnica es ampliamente utilizada en diversos campos como la ciencia de datos, el aprendizaje automático y la inteligencia artificial.
- Se utilizan para resolver problemas de **aprendizaje supervisado**, donde se utiliza un **conjunto de datos** (*datasets*) con ejemplos etiquetados para entrenar el modelo.

Modelos de clasificación - funcionamiento

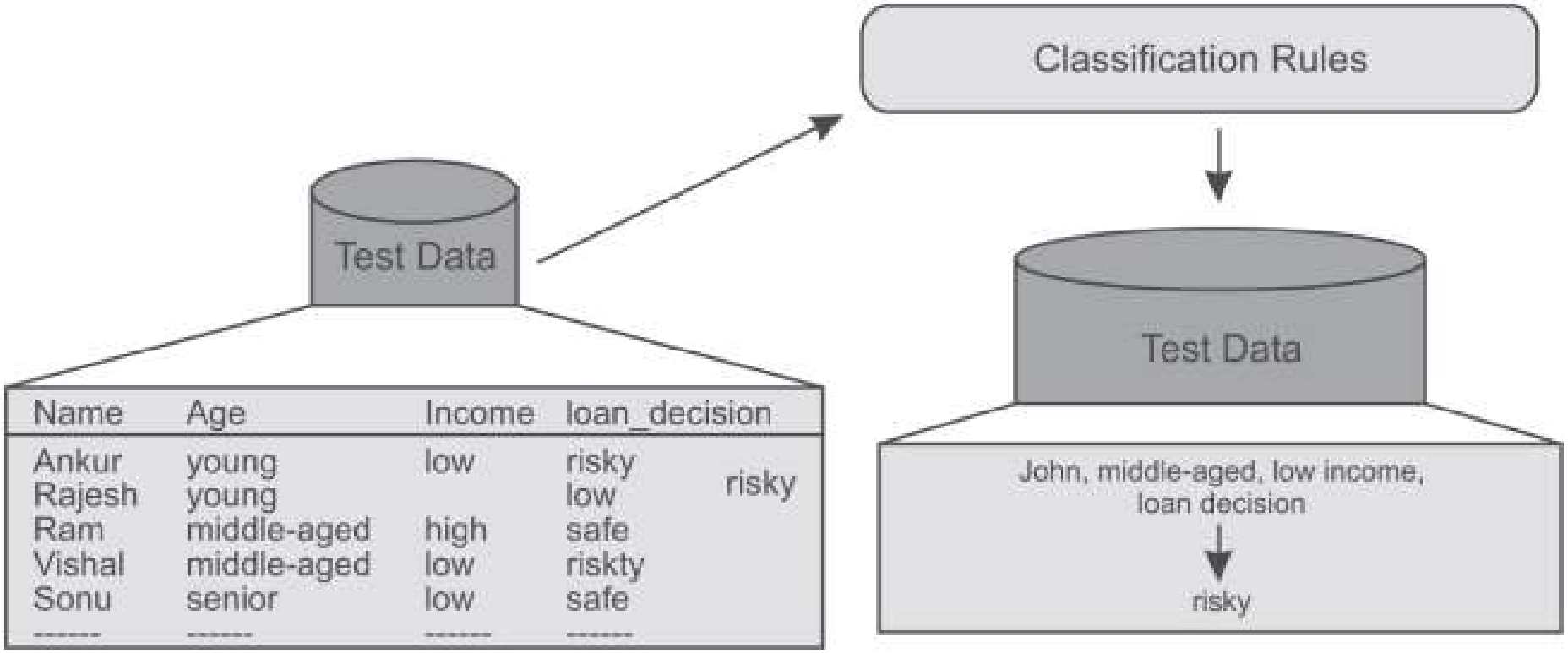
- **Entrenamiento**: el modelo se entrena con un **dataset** etiquetados, llamado **conjunto de entrenamiento**.
- Cada ejemplo del conjunto de entrenamiento tiene un conjunto de **atributos** (**características**) y una **clase** (**etiqueta**).
- **Predicción**: después del entrenamiento, el modelo puede predecir la clase de nuevos ejemplos que no se utilizaron en el entrenamiento.
- Tipos de problemas de clasificación:
  - **Binario**: clasificar en dos clases (por ejemplo, spam/no spam, positivo/negativo)
  - **Multiclase**: clasificar en más de dos clases (por ejemplo, tipo de flor, tipo de cáncer)

Modelos de clasificación - entrenamiento



Fuente: Bhatia, P. Data Mining and Data Warehousing: Principles and Practical Techniques. Cambridge University Press, 2019

Modelos de clasificación - Predicción



Fuente: Bhatia, P. Data Mining and Data Warehousing: Principles and Practical Techniques. Cambridge University Press, 2019

Tipos de modelos de clasificación

- **Regresión logística**: **Modelo probabilístico** que utiliza la **función logística** para estimar la probabilidad de que un ejemplo pertenezca a una clase.
- **Árboles de decisión**: **estructura jerárquica** que divide el espacio de características en subconjuntos, creando **reglas de decisión** para clasificar ejemplos.
- **KNN** (K-Vecinos más cercanos): Algoritmo que clasifica un ejemplo en función de la clase de los K ejemplos más cercanos a él en el espacio de características.
  - **Clasificación por similaridad**
  - **Algoritmo simple y eficiente**
  - Elección de **K** y métricas de distancia.

Tipos de modelos de clasificación … cont.

- **Naive Bayes**: **Modelo probabilístico ingenuo** que asume independencia entre atributos, utilizando el **teorema de Bayes** para calcular la probabilidad de que un ejemplo pertenezca a una clase.
- **Redes Neuronales Artificiales**: Arquitecturas computacionales inspiradas en el cerebro humano, capaces de aprender relaciones complejas entre atributos.
  - **Arquitectura de redes neurales**
  - **Perceptron**, MLP, CNN, RNN
  - **Entrenamiento** y **optimización**
  - Aplicaciones en clasificación de **imágenes,** texto, etc.

Tipos de modelos de clasificación …cont.

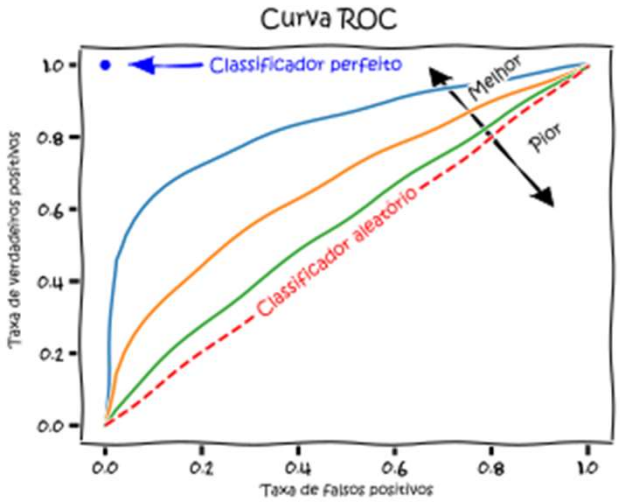
- **Máquinas de vectores de soporte (SVM)**: Algoritmo que busca encontrar el **hiperplano de soporte máximo** que separa los ejemplos de diferentes clases.
  - Función del núcleo (kernel)
  - Margen suave para manejar datos no linealmente separables.
- **Aprendizaje ensemble**: combinación de diferentes modelos para mejorar el rendimiento general de la clasificación.

Métricas de evaluación

- **Acuracy**: Proporción de ejemplos clasificados correctamente.
- **Precisión**: Proporción de ejemplos que fueron clasificados como positivos y que realmente lo son.
- **Recall**: Proporción de ejemplos que realmente son positivos y fueron clasificados como positivos.
- **F1-score**: media armónica entre precisión y recuperación.
- **Curva ROC**: Curva que muestra la relación entre la tasa de verdaderos positivos y la tasa de falsos positivos.
- **Matriz de confusión**: tabla que muestra el recuento de ejemplos para cada combinación de clase real y clase prevista.

Métricas de evaluación

- acuracy $= \frac{predicciones\ correctas}{total\ de\ predicciones} = \frac{VP+VN}{VP+VN+FP+F}$
- Precisión$=\frac{Predicciones\ positivas\ correctas}{Predicciones\ positivas} = \frac{VP}{VP+FP}$
- Recall $= \frac{VP}{VP+FN}$
- $F1-score = \frac{2}{Recall^{-1}+Precisión^{-1}} = 2 * \frac{Precisión*Recall}{Precisión+Recall}$



	N	P
N	VN	FP
P	FN	VP

Matriz de confusión

Consideraciones importantes

- **Elección del modelo**: Depende del tipo de problema, datos disponibles y recursos computacionales.
- **Preprocesamiento de datos**: Limpiar, transformar y normalizar los datos es fundamental para el buen rendimiento de los modelos.
- **Interpretabilidad vs. Rendimiento**: los modelos más complejos pueden tener un mejor rendimiento, pero ser menos interpretables.
- **Monitoreo y actualización de modelos**: es importante monitorear el desempeño de los modelos a lo largo del tiempo y volver a entrenarlos cuando sea necesario.

Conclusiones

- Los **modelos de clasificación** son herramientas poderosas para analizar e interpretar datos, con aplicaciones en varias áreas, como:
  - **Marketing**: segmentar clientes, predecir la deserción.
  - **Salud**: diagnosticar enfermedades, predecir el riesgo de recurrencia.
  - **Finanzas**: detectar fraude, predecir el riesgo crediticio.
  - **Otras áreas**: industria, manufactura, agricultura, etc.
- Elegir el modelo ideal, preprocesar los datos y evaluar el desempeño son pasos importantes para garantizar la calidad de los resultados.