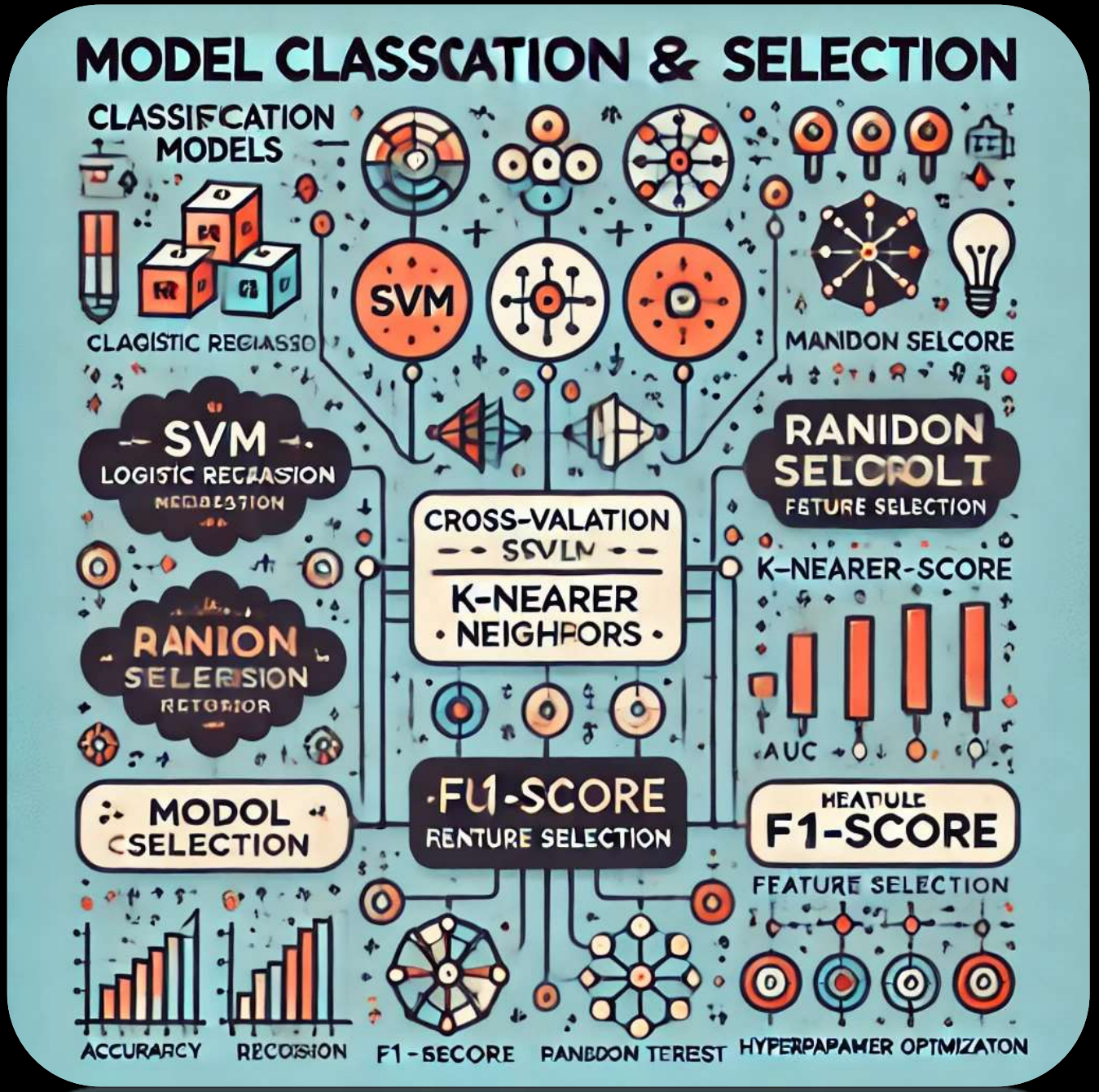

MODEL CLASSCATION & SELECTION
CLASSIFICATION MODELS
SVM
SVM
LOGISTIC REGRESSION
RANDOM SELECTION
FEATURE SELECTION
CROSS-VALIDATION
SCORE
K-NEAREST NEIGHBORS
K-NEAREST-SCORE
AUC
MODEL SELECTION
F1-SCORE
FEATURE SELECTION
F1-SCORE
FEATURE SELECTION
ACCURACY
RECISION
F1-SCORE
RANDOM FOREST
HYPERPARAMETER OPTIMIZATION

Clasificación y Selección de Modelos

La clasificación y selección de modelos son etapas cruciales en el proceso de desarrollo de modelos de aprendizaje automático. La **clasificación** se refiere a la tarea de predecir una categoría para una muestra de datos, mientras que la **selección de modelos** implica elegir el mejor modelo entre una variedad de candidatos, basado en su rendimiento.

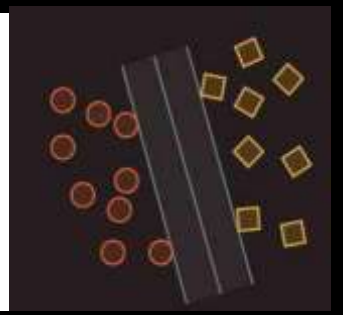
Métricas de Evaluación y Regularización

Clasificación

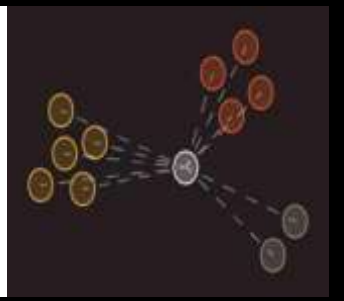
La clasificación es una técnica de aprendizaje supervisado donde el objetivo es predecir la etiqueta de clase de las muestras con base en uno o más características de entrada. Algunos algoritmos comunes de clasificación incluyen:

1. **Regresión Logística:**
   - Modelo estadístico que predice la probabilidad de una categoría.
   - Útil para problemas de clasificación binaria.

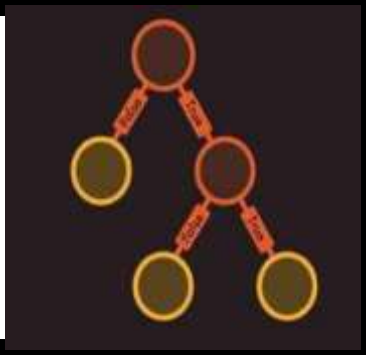
2. **Máquinas de Soporte Vectorial (SVM):**
   - Encuentra el hiperplano que mejor separa las clases en el espacio de características.
   - Eficaz en espacios de alta dimensionalidad.



3. **K-Nearest Neighbors (KNN):**
   - Clasifica una muestra basándose en la mayoría de las clases de sus vecinos más cercanos.
   - Simple pero puede ser computacionalmente costoso.



4. **Árboles de Decisión**:
    - Modelo basado en reglas que se divide en ramas para hacer predicciones.
    - Fácil de interpretar pero propenso al sobreajuste.



5. **Random Forest**:
    - Combina múltiples árboles de decisión para mejorar la precisión y controlar el sobreajuste.
    - Robusto y eficaz para muchos tipos de datos.

6. **Redes Neuronales Artificiales**:
    - Modelos inspirados en el cerebro humano que son potentes para detectar patrones complejos.
    - Requieren una gran cantidad de datos y potencia computacional.

Ventajas

•**Simplicidad**: Los modelos son más simples y fáciles de interpretar.

•**Rapidez**: Requieren menos datos y son más rápidos de entrenar.

Desventajas

•**Rigidez**: Si las suposiciones sobre la distribución de los datos son incorrectas, el modelo puede no funcionar bien.

Ventajas

•Flexibilidad: Pueden adaptarse mejor a la estructura real de los datos.

•Menos Suposiciones: No requieren suposiciones estrictas sobre la distribución de los datos.

Desventajas

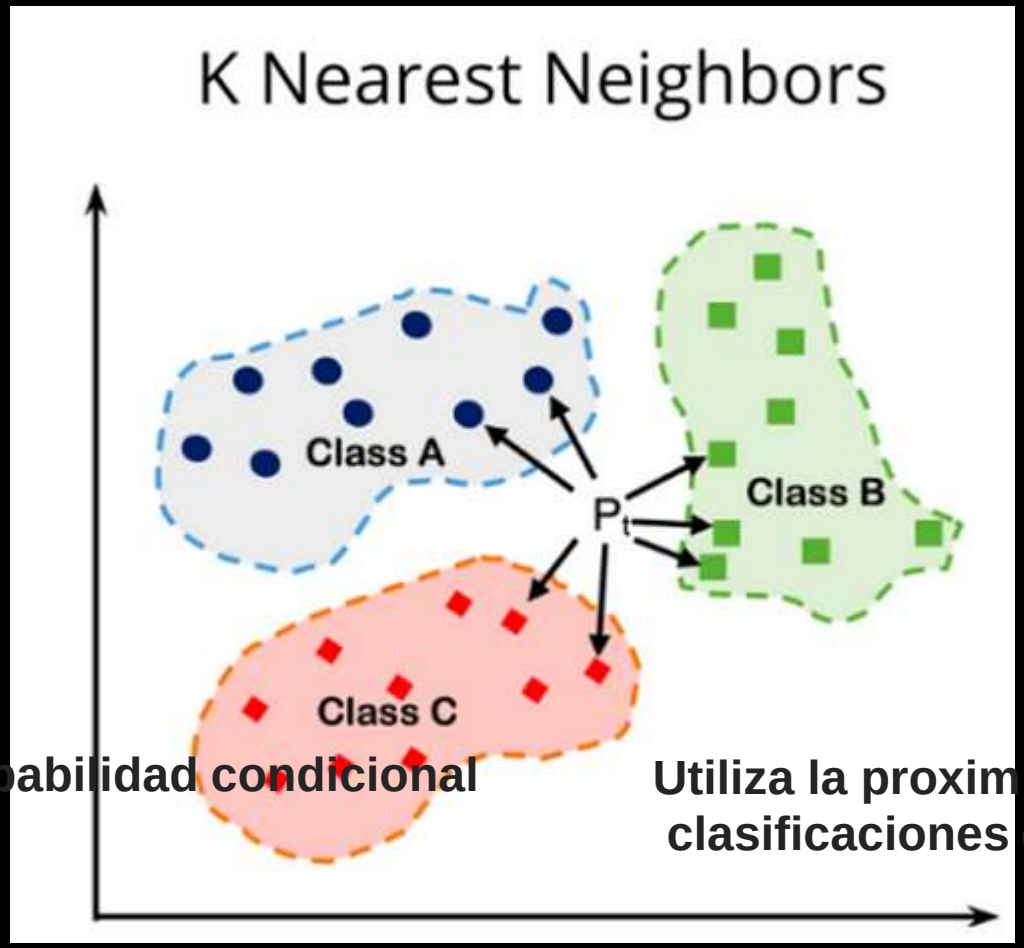
•Complejidad: Pueden ser más difíciles de interpretar.

•Computacionalmente Intensivos: Requieren más datos y tiempo para entrenar.

Comparación

Característica	Métodos Paramétricos	Métodos No Paramétricos
Suposiciones	Fuertes	Pocas
Número de Parámetros	Fijo	Crece con los datos
Flexibilidad	Menos flexible	Más flexible
Requisitos de Datos	Menos datos	Más datos
Interpretabilidad	Más interpretables	Menos interpretables
Ejemplos	Regresión Lineal, Regresión Logística	KNN, Árboles de Decisión, Random Forest

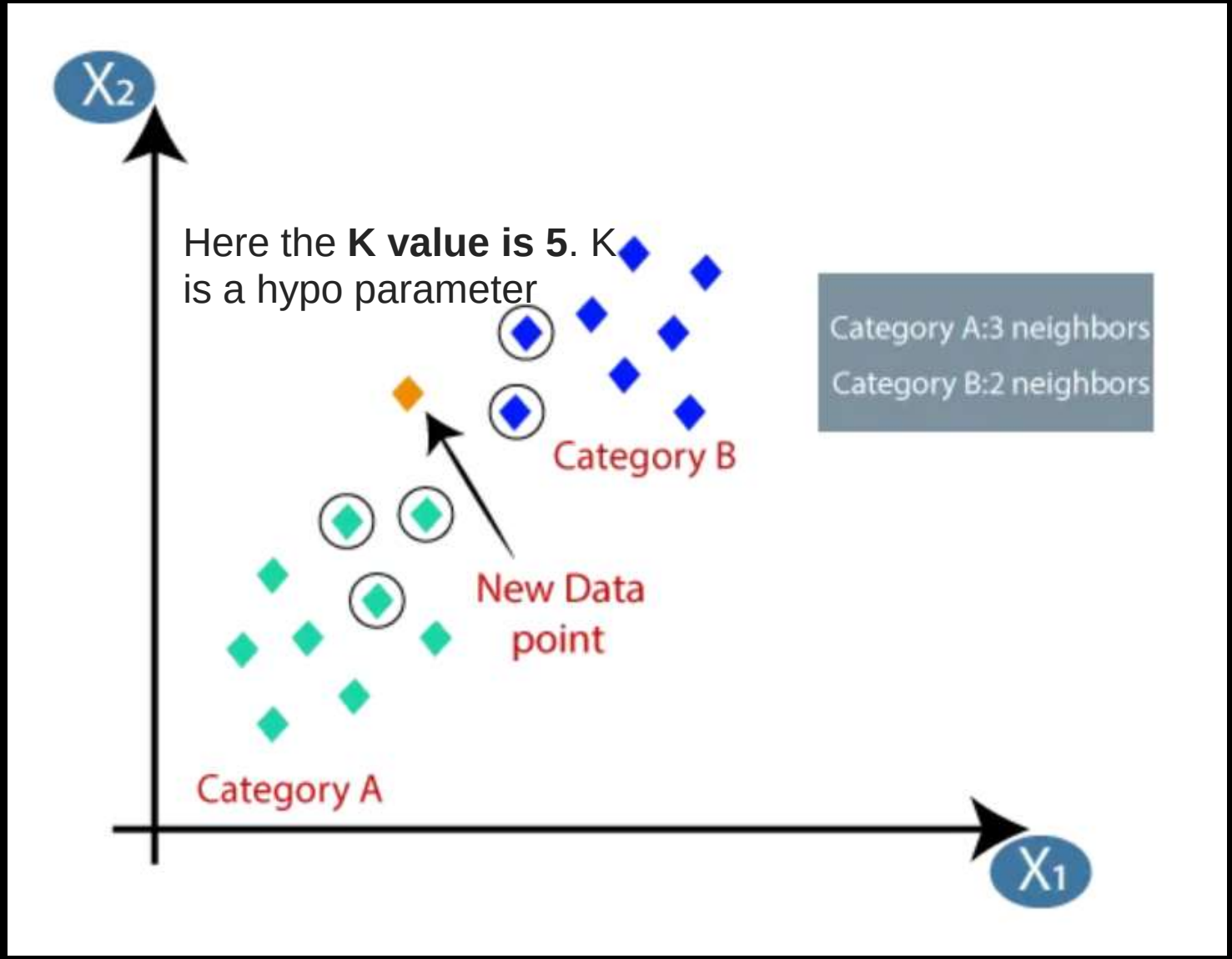
Algoritmo de los K vecinos más cercanos



Determinar la probabilidad condicional

Utiliza la proximidad para hacer clasificaciones o predicciones

Algoritmo de los K vecinos más próximos (KNN)



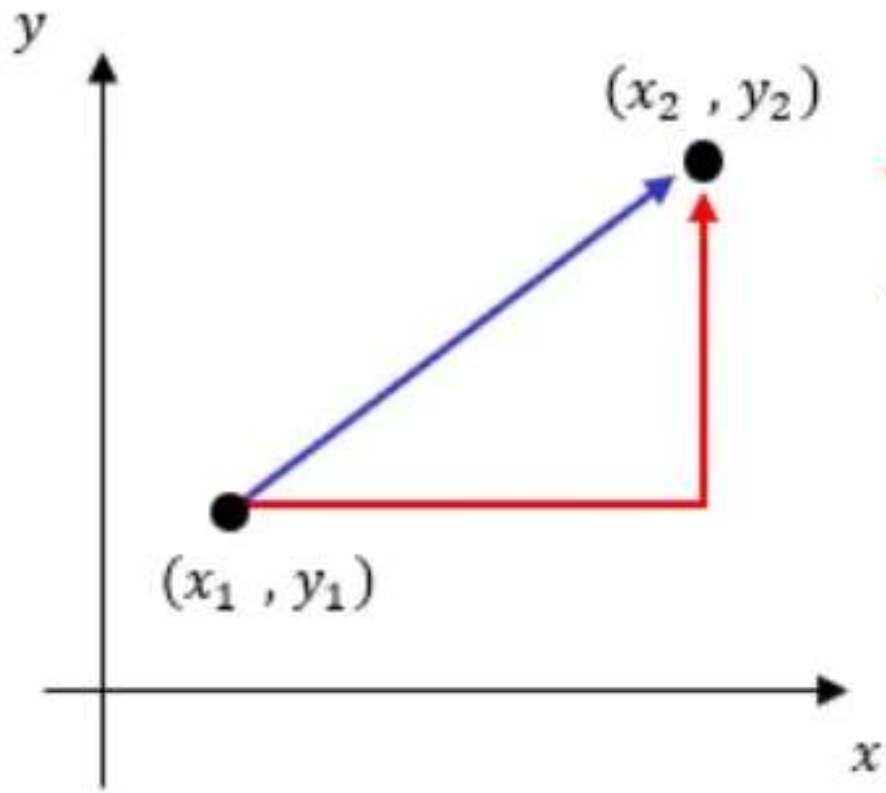
- **distancia euclidiana**
- **Distancia de Manhattan**

Para calcular la distancia euclidiana entre dos puntos, denotada por

```
d = √[ (x2- x1)^ 2 + (y2- y1)^ 2 ].
```

Por otro lado, para la distancia de Manhattan, calculamos la distancia entre dos puntos.

```
d = |x1 - x2| + |y1 - y2|
```

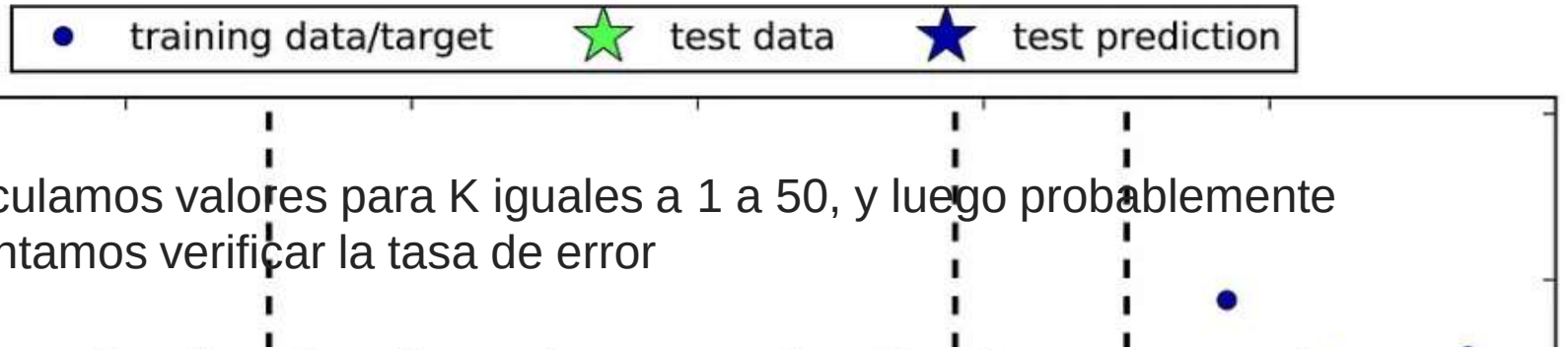


Manhattan Distance L^1

Euclidean Distance L^2

$$L^1 = |x_2 - x_1| + |y_2 - y_1|$$

$$L^2 = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

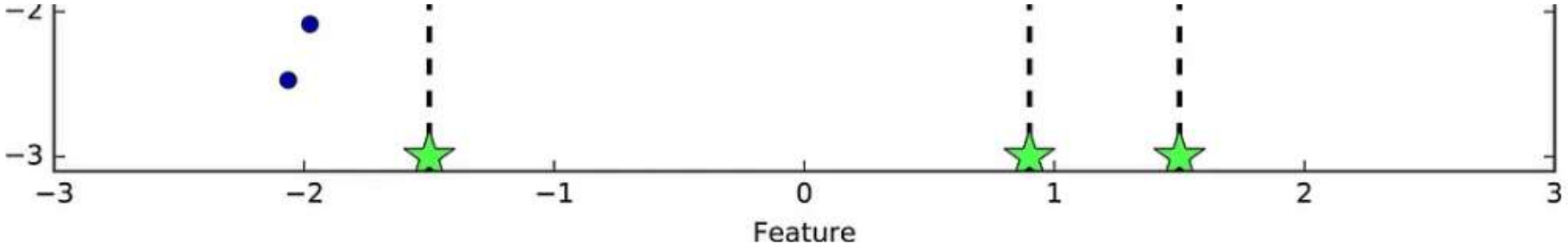


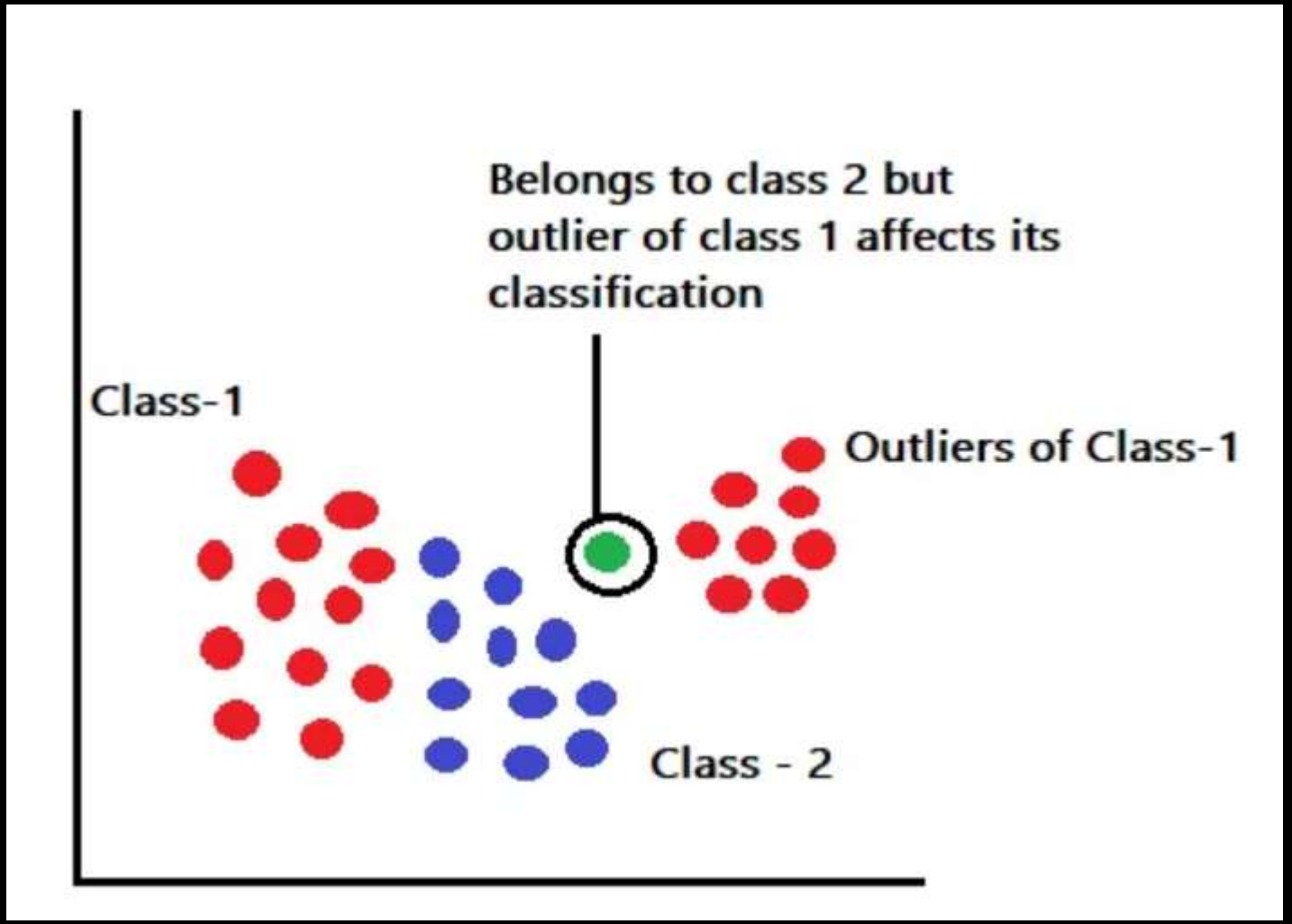
Calculamos valores para K iguales a 1 a 50, y luego probablemente intentamos verificar la tasa de error

Dos cosas más sobre el vecino más cercano de K funcionan muy mal

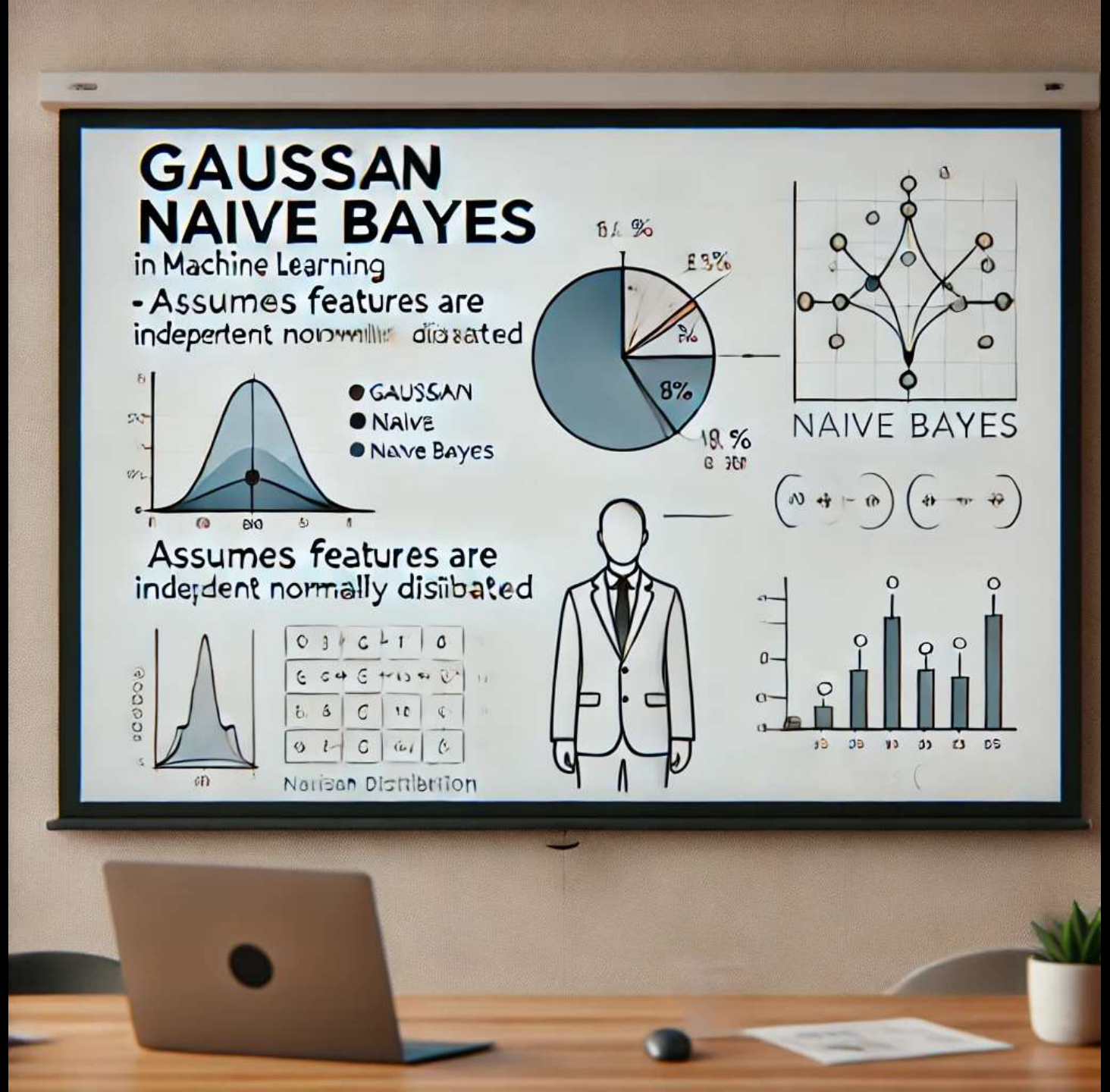
- Valores atípicos
- Conjunto de datos desequilibrado

Entonces, la limitación de KNN es que, debido a la presencia de un conjunto de datos desequilibrado o atípico, podría clasificar erróneamente los puntos.




Belongs to class 2 but
outlier of class 1 affects its
classification
Class-1
Outliers of Class-1
Class - 2

- **Escalabilidad**: Puede ser lento para conjuntos de datos grandes porque requiere calcular la distancia a todos los puntos en el conjunto de entrenamiento.
- **Sensibilidad a la Escala de los Datos**: Las características con escalas mayores pueden dominar la distancia.
- **Almacenamiento**: Necesita almacenar todos los datos de entrenamiento.

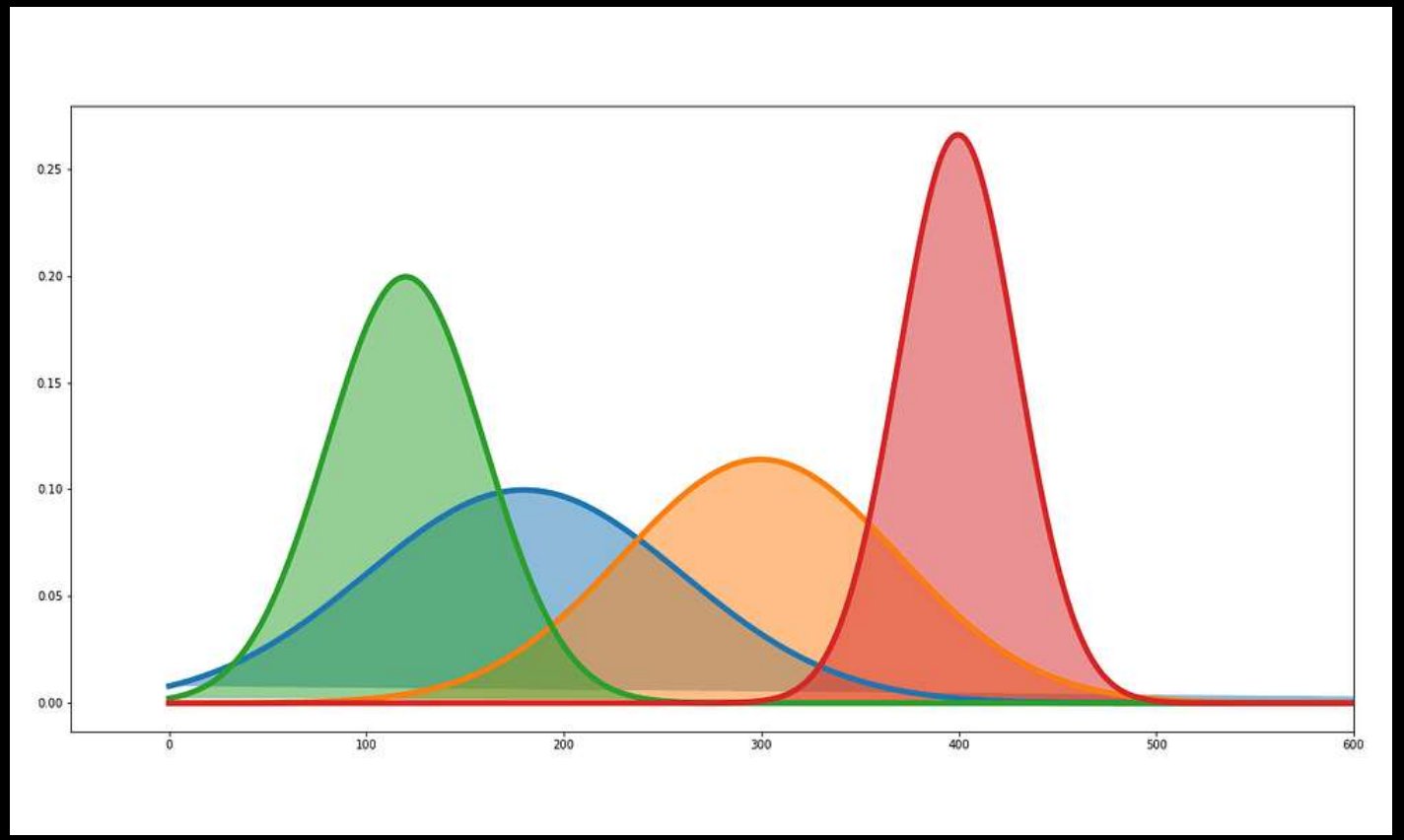

GAUSSAN
NAIVE BAYES
in Machine Learning
- Assumes features are
independent normally distributed
GAUSSAN
Naive
Naive Bayes
8%
NAIVE BAYES
Assumes features are
independent normally distributed

Método Bayes Gaussiano

Distribución Normal

La distribución normal es un patrón estadístico que aparece cuando un conjunto de datos se distribuye de manera uniforme alrededor de un valor central.

Es decir, que la mayoría de las observaciones se agrupan en torno al promedio, y los valores se vuelven progresivamente menos comunes a medida que se alejan de este punto medio.

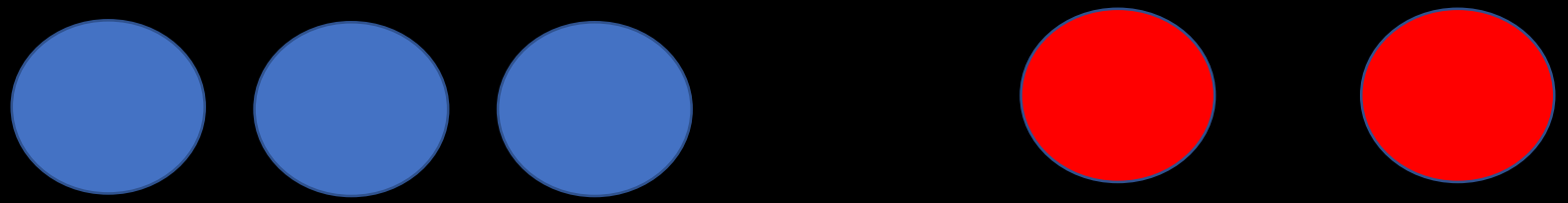


El teorema calcula la probabilidad posterior, que es la probabilidad de la hipótesis (la fruta es una manzana) dados nuestros datos observados (color, tamaño, sabor).

El algoritmo Bayes ingenuo gaussiano, un campeón entre los algoritmos de clasificación, se destaca en situaciones del mundo real donde los puntos de datos son continuos y se puede suponer que siguen una distribución normal. Aprovecha el principio del teorema de Bayes para hacer predicciones considerando las distribuciones de características condicionales. Este método implica datos de entrenamiento para estimar los parámetros de la distribución normal, específicamente la media y la desviación estándar, para cada característica dentro de cada clase.

Los fundamentos matemáticos del método Bayes ingenuo

La teoría que lo sustenta es bastante sencilla y se centra en la probabilidad de que ocurra un evento dada la presencia de otro evento. Es decir consiste en preguntar cómo cambia la probabilidad de un evento cuando ya sabemos que algo ha sucedido



la probabilidad condicional también ayuda a evaluar las relaciones entre las variables de un conjunto de datos. Saber que "B ya ocurrió" puede cambiar drásticamente la "probabilidad condicional de A dado B", que es la base para hacer predicciones con Naive Bayes. Esta interacción de probabilidades es lo que permite que el modelo navegue por los conjuntos de datos y haga predicciones fundamentadas basadas en la evidencia proporcionada.

Gaussiano

Cuando se utiliza el clasificador bayesiano ingenuo gaussiano, se hace la siguiente suposición:

- Los valores continuos de cada característica se distribuyen de manera normal dentro de cada clase.

Esto significa que, para cada característica x y cada clase y, la distribución de x dado y sigue una distribución normal con una media μ_y y una desviación estándar σ_y.

$$P(x \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x-\mu_y)^2}{2\sigma_y^2}\right)$$

Fórmula

La probabilidad de una característica x dado una clase y se puede calcular utilizando la función de densidad de probabilidad de la distribución normal:

$$P(x \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x-\mu_y)^2}{2\sigma_y^2}\right)$$

Proceso de Clasificación

1. **Estimación de Parámetros**: Para cada clase y, se estiman los parámetros μ_y y σ_y de la distribución normal utilizando los datos de entrenamiento.
2. **Cálculo de Probabilidades Posteriores**: Para una instancia nueva x, se calcula la probabilidad posterior de cada clase y:

$$P(y \mid x) \propto P(y) \prod_{i=1}^{n} P(x_i \mid y)$$

Donde $P(y)$ es la probabilidad previa de la clase y, y $P(x_i \mid y)$ es la probabilidad de la característica x_i dado la clase y.

3. **Clasificación**: La clase y con la probabilidad posterior más alta es la predicción del clasificador.

Resumen de Diferencias

- **Bayes Ingenuo Gaussiano**: Basado en el teorema de Bayes, asume independencia y normalidad, eficiente en entrenamiento y predicción, menos sensible a ruidos.
- **K Vecinos Más Cercanos (KNN)**: Basado en distancias y vecinos cercanos, no asume distribuciones, costoso en predicción, más flexible y aplicable a diversos problemas.

Estos puntos resumen las diferencias fundamentales entre ambos métodos, ofreciendo una perspectiva clara de cuándo podría ser más adecuado usar uno sobre el otro.

EJEMPLO APLICADO

Explicación:

- K-Nearest Neighbors (K-NN):
  - **Ajuste de `n_neighbors`**: Al aumentar el número de vecinos a 15, K-NN utiliza más información de los vecinos cercanos para clasificar cada punto de datos. Esto a menudo mejora la estabilidad de las predicciones, especialmente en conjuntos de datos ruidosos o en problemas con muchos ejemplos similares.
  - **Resultado**: Con `n_neighbors=15`, el modelo K-NN ahora es más preciso que el modelo Bayes Ingenuo Gaussiano en este conjunto de datos. Esto ocurre porque, en este caso, la decisión de promediar sobre más vecinos proporciona una clasificación más robusta.
- Bayes Ingenuo Gaussiano:
  - **Suposiciones**: Este modelo asume que las características siguen una distribución normal (gaussiana) y que las características son independientes entre sí. Aunque este modelo es rápido y eficiente, sus suposiciones pueden no capturar completamente la complejidad del conjunto de datos del cáncer de mama.
  - **Resultado**: La precisión del 94.15% sugiere que el modelo es bastante bueno, pero en este caso específico, K-NN con más vecinos logró superar su rendimiento.