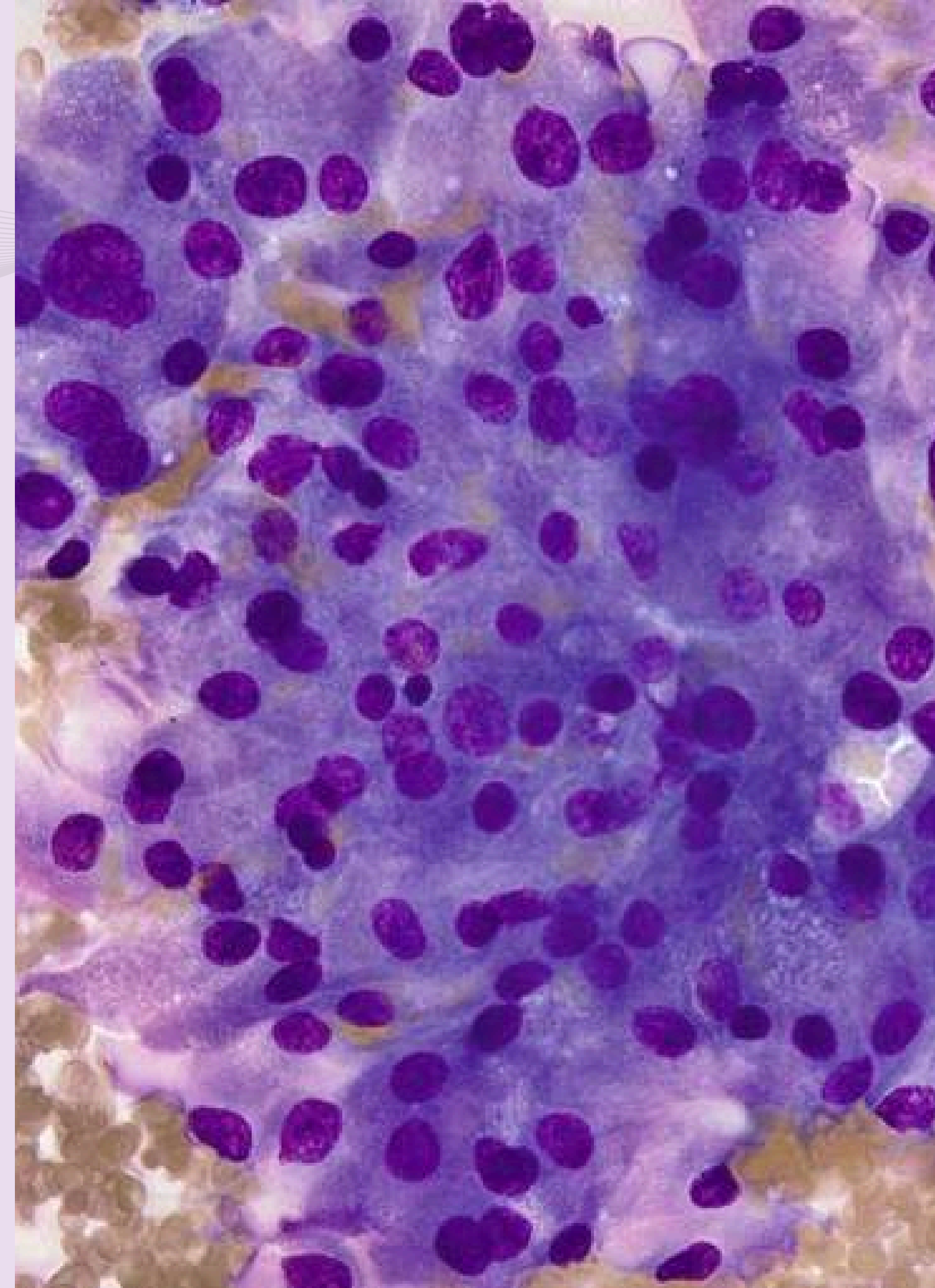


**Fundamentos de
Ingeniería de Datos**

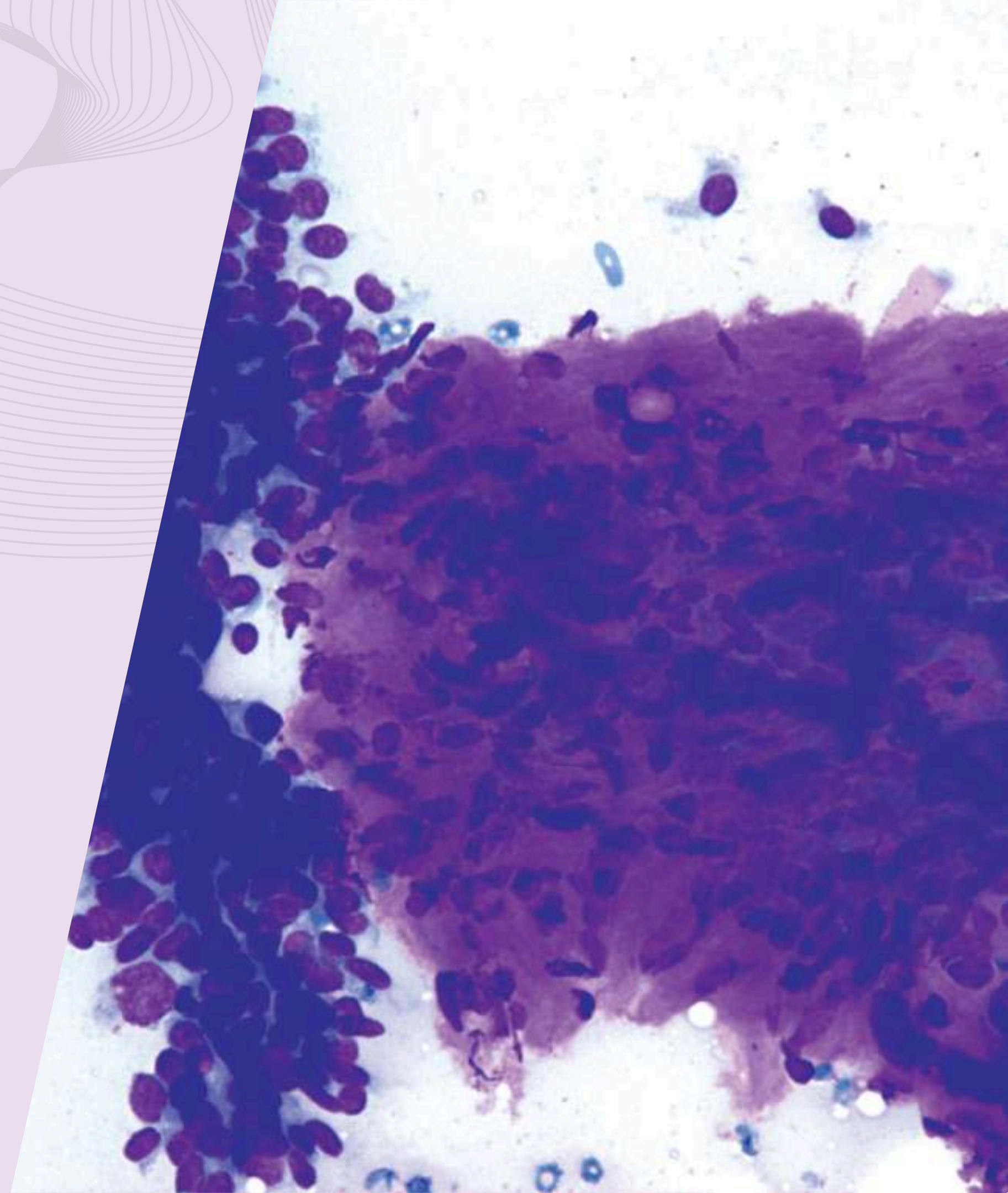
Análisis de Datos de Imágenes FNA para el Diagnóstico de Cáncer de Mama

Claudia Teresa Heredia Ceballos
Javier Fernández Castillo
Manuel Otero Barbasán
Marta Pineda Gisbert



Contenido

- **Introducción**
- **Metodología**
- **Implementación**
- **Resultados**
- **Conclusiones**

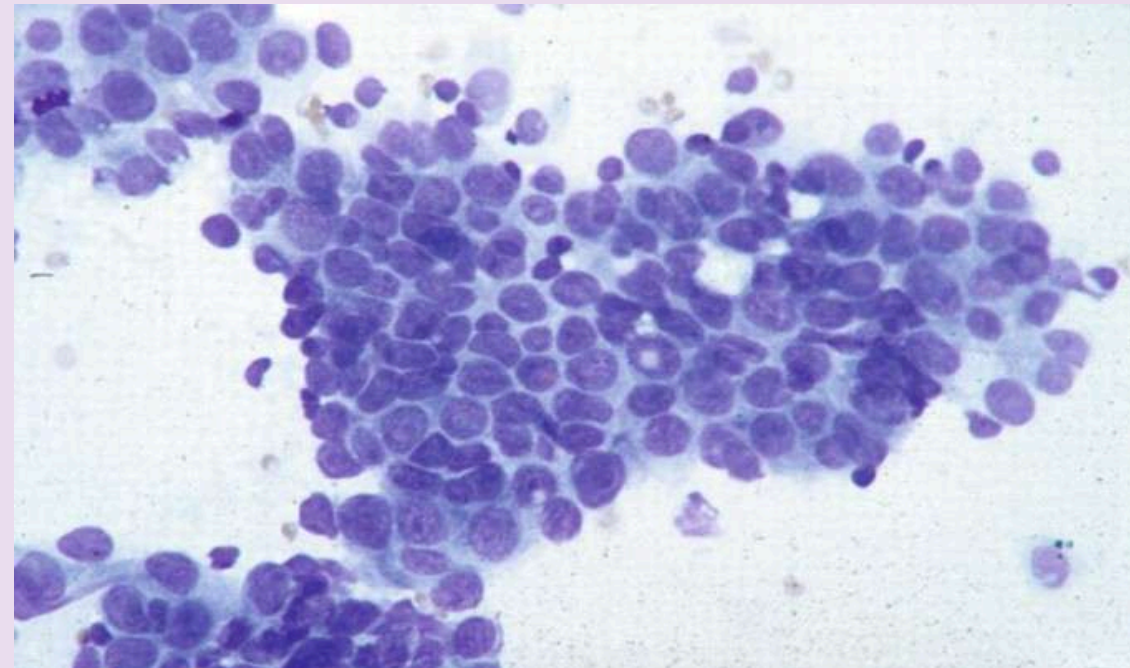
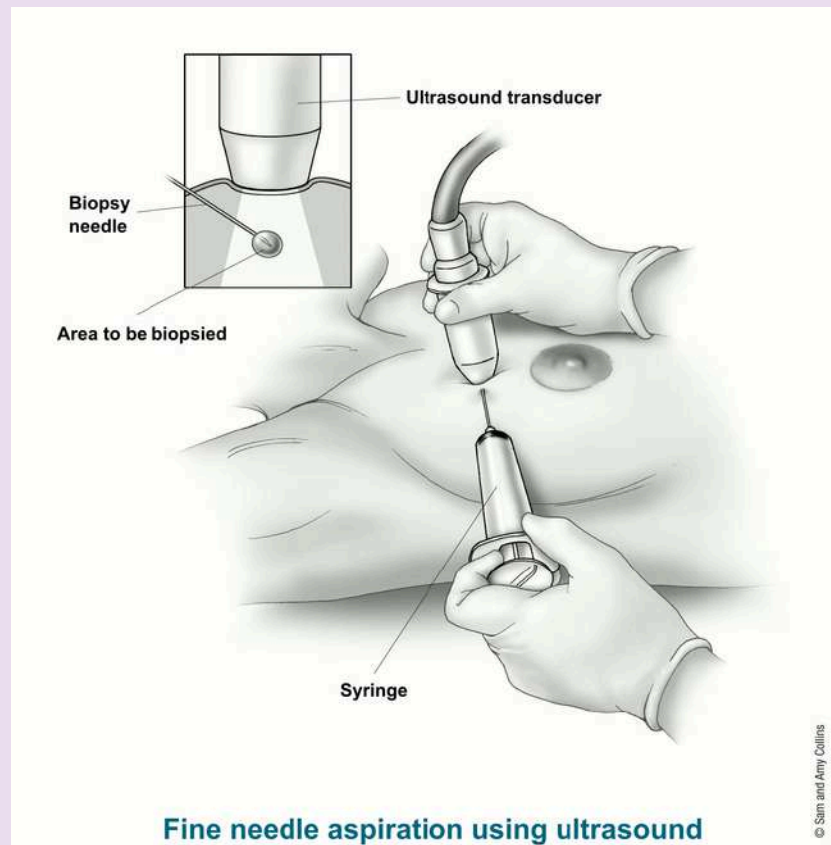


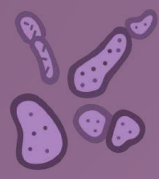


INTRODUCCIÓN



Predicción del diagnóstico de cáncer de mama a través de Imágenes FNA



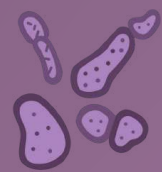


Hospital de Wisconsin 1993





METODOLOGÍA



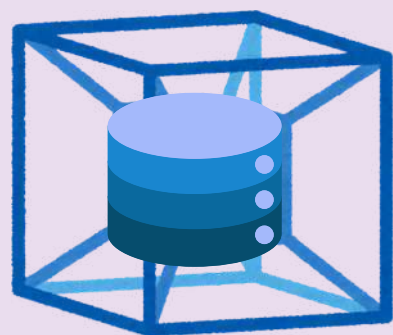
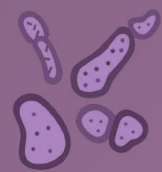
Roadmap

7





IMPLEMENTACIÓN



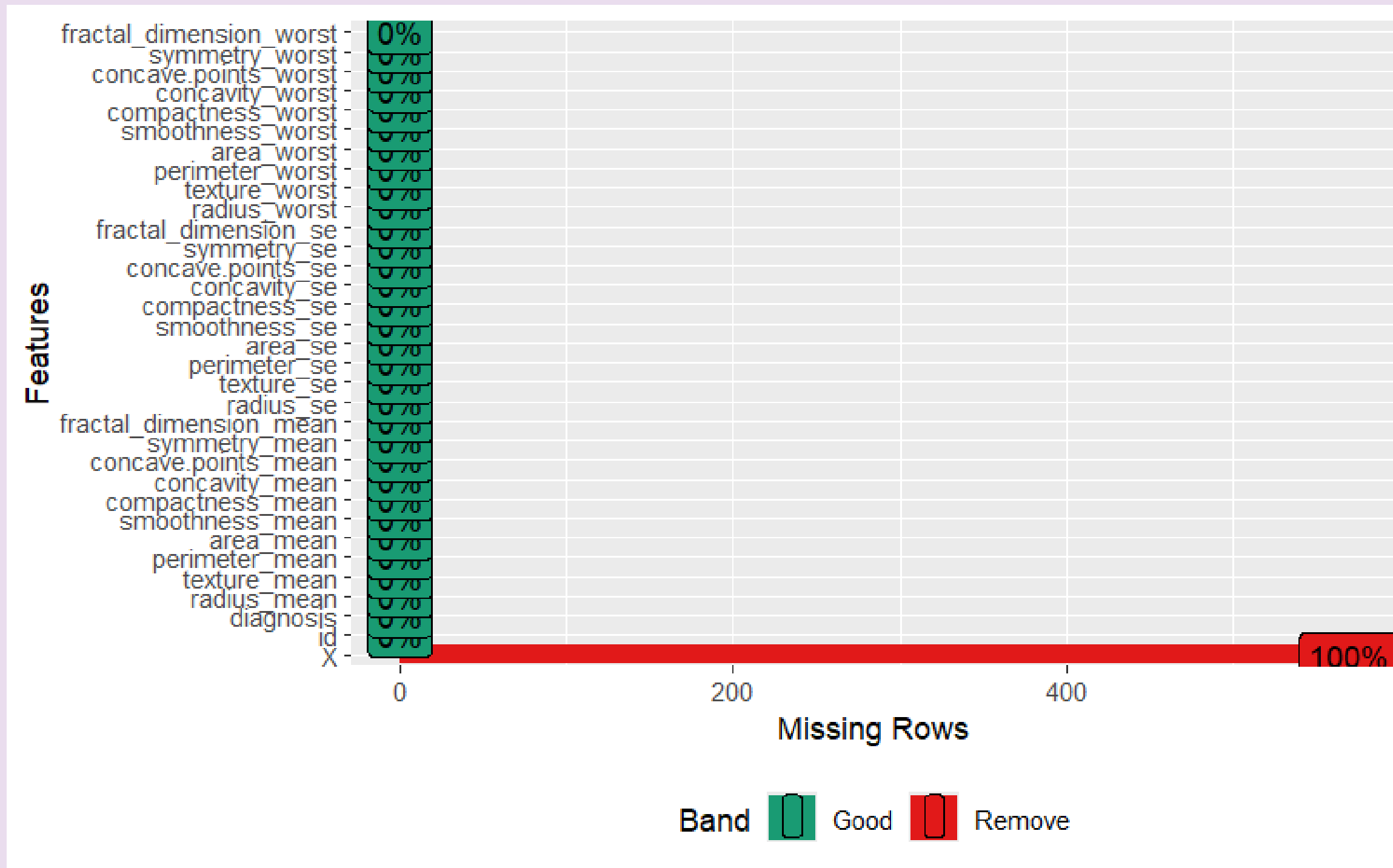
33 columnas

569 filas

31 variables numéricas
1 variable entera (ID)

1 variable categórica
VALOR OBJETIVO

- **ID (1), X(1)**
- **DIAGNÓSTICO (1)**
- **RADIO (3), PERÍMETRO (3), ÁREA (3),**
- **COMPACIDAD(3), SUAVIDAD(3),**
- **CONCAVIDAD (3), PTS. CONCAVOS (3), SIMETRÍA (3),**
- **TEXTURA (3), DIMENSIÓN FRACTAL (1)**



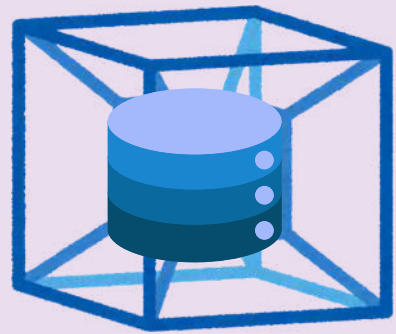
Eliminación de variables:

- ID
- X



Análisis y preprocesamiento de los datos

II



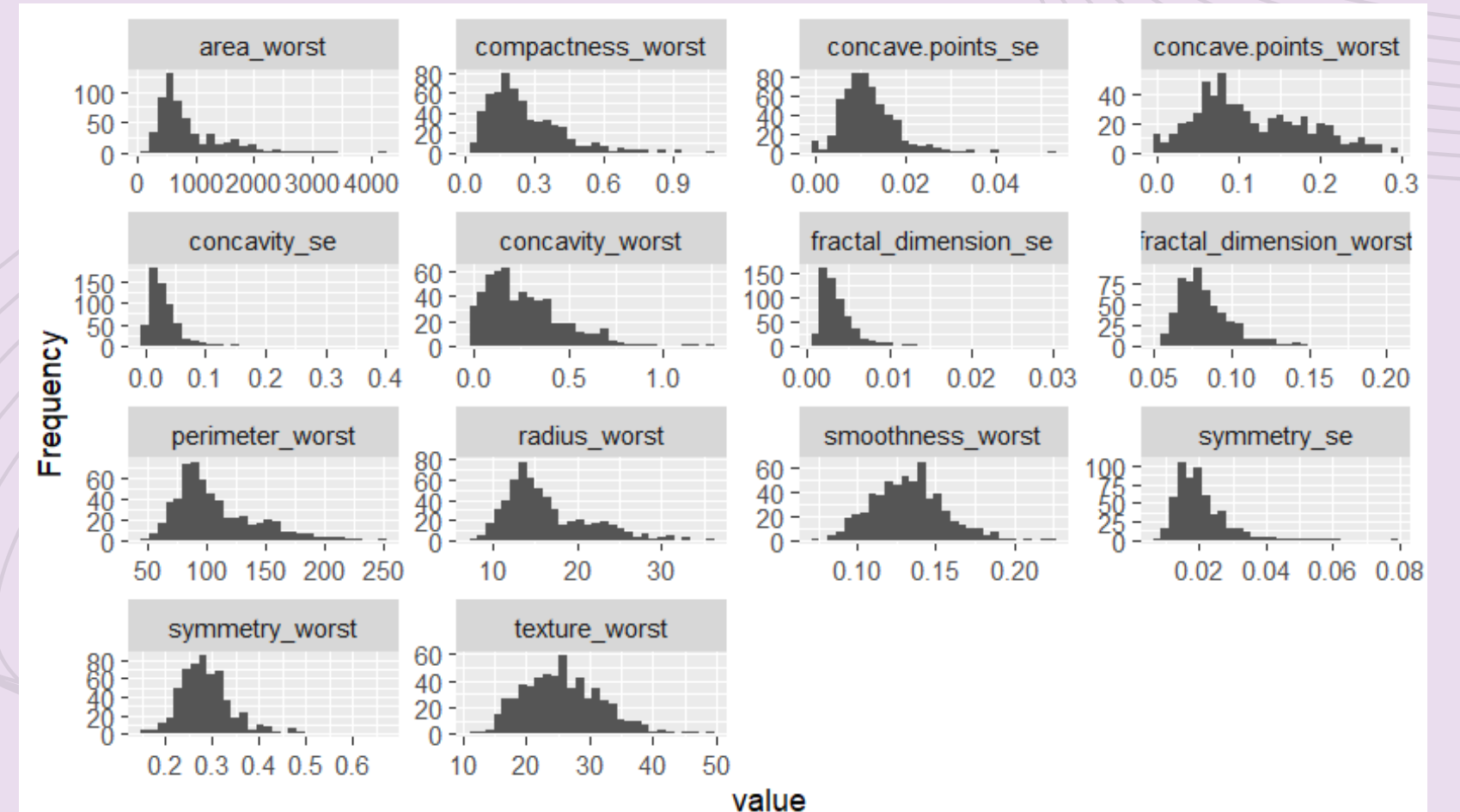
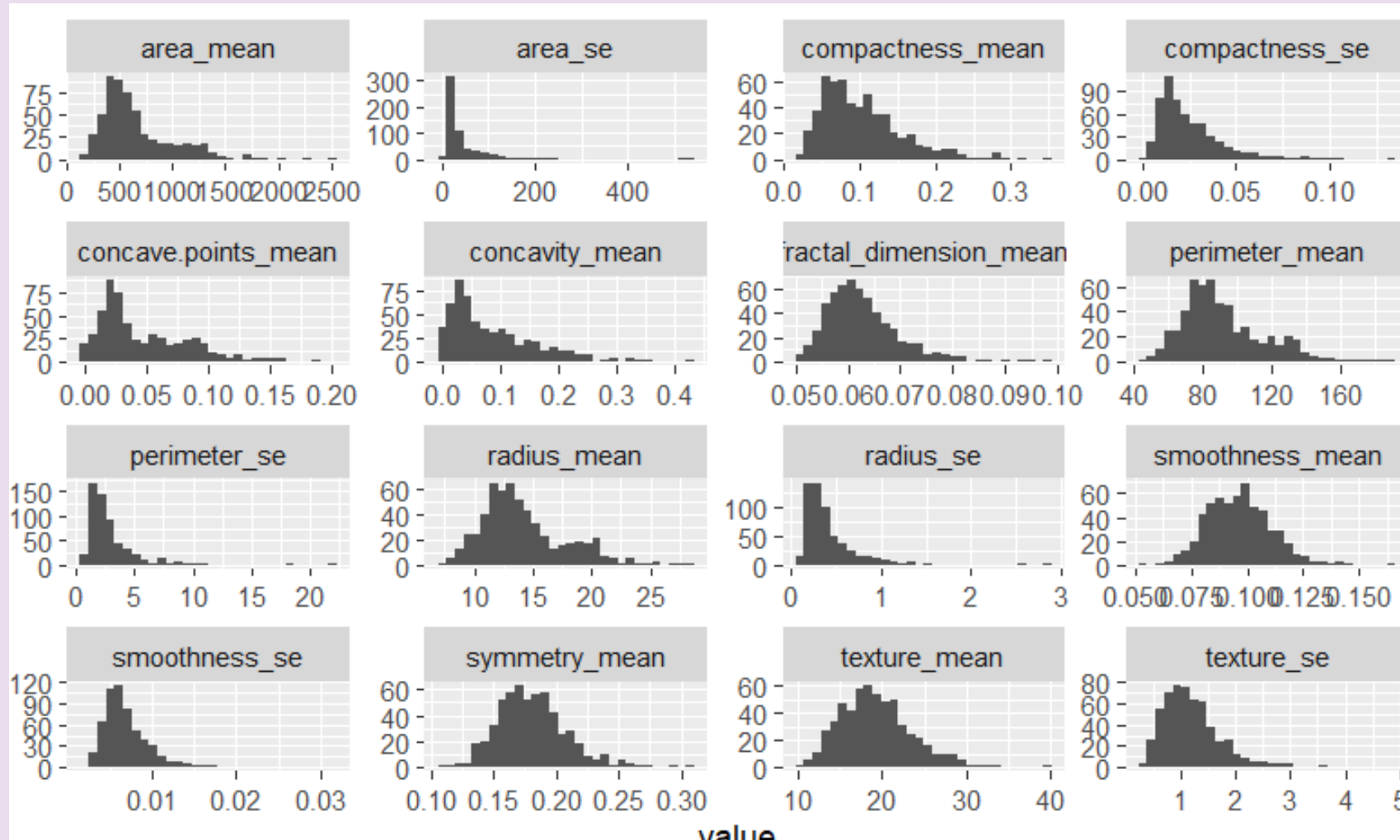
31 columnas

569 filas

30 variables numéricas

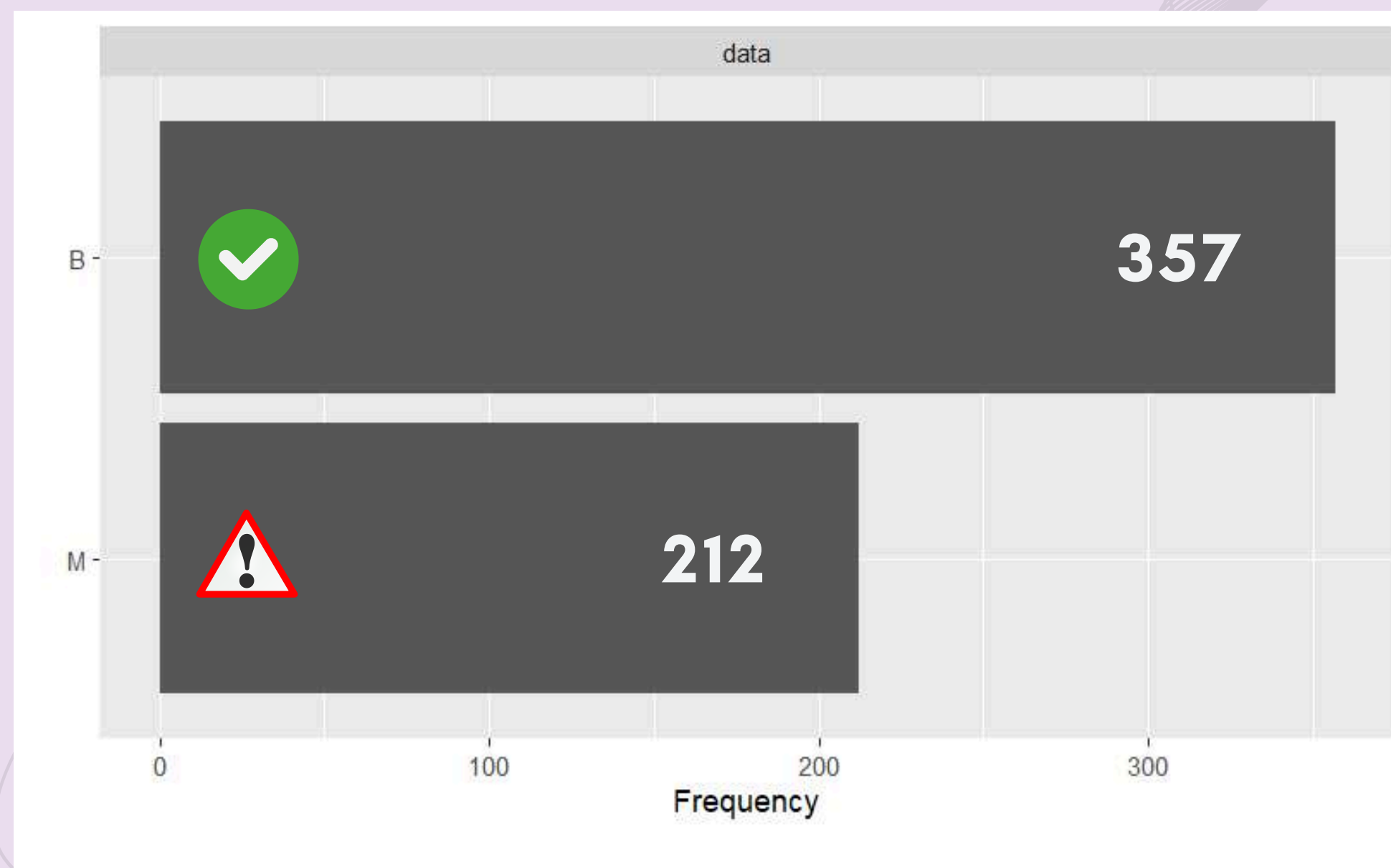
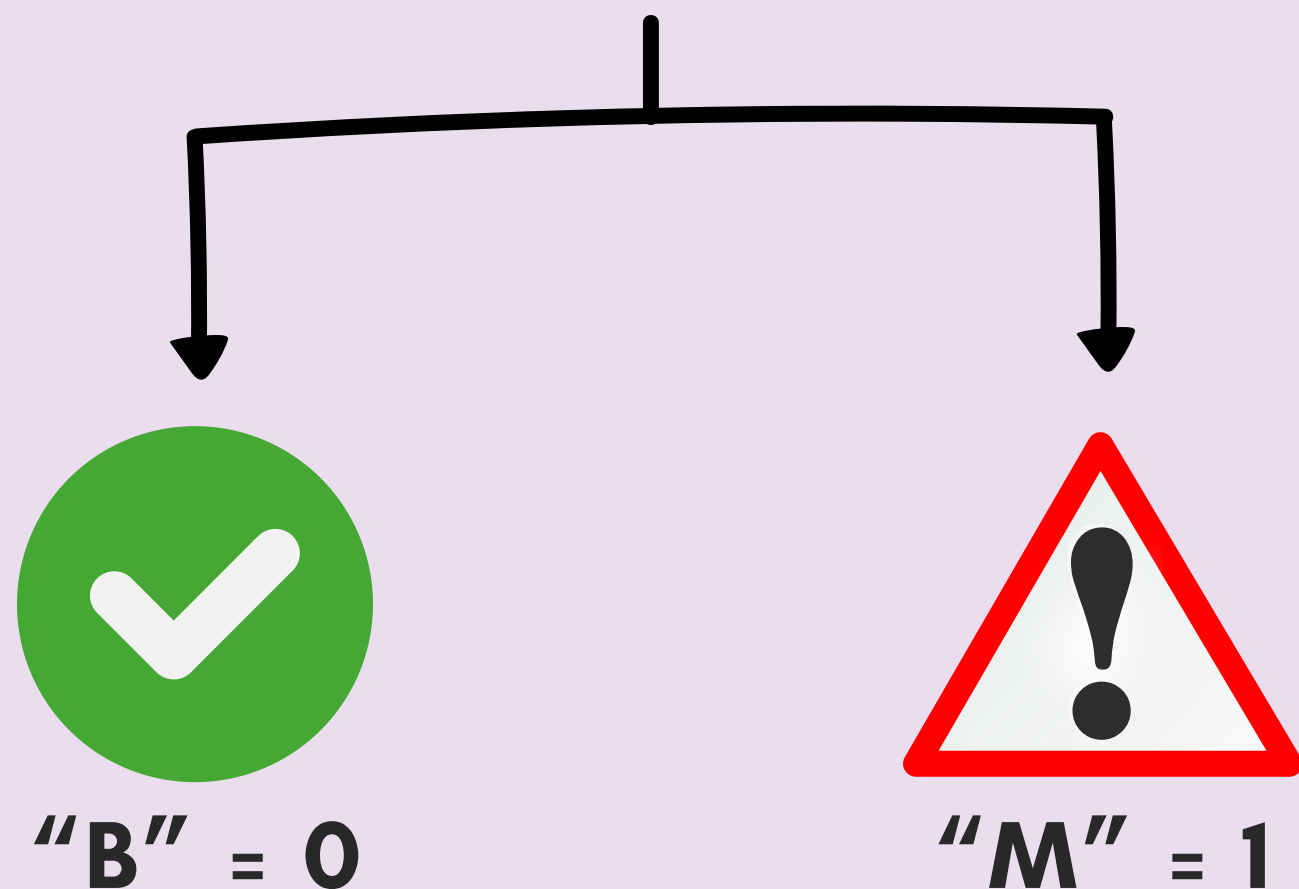
1 variable categórica

VALOR OBJETIVO



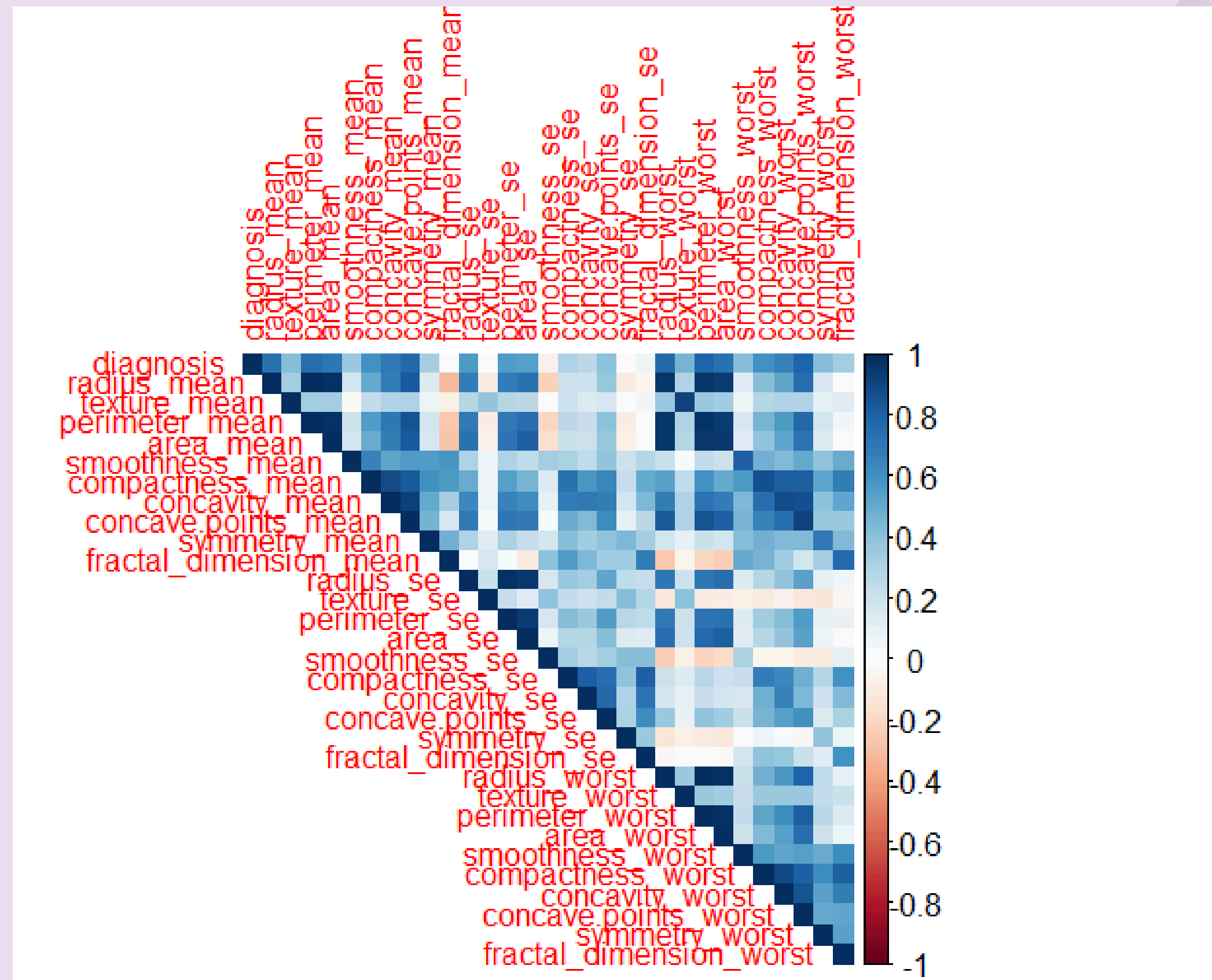


Diagnosis





Estudio de correlación entre variables





Estudio de correlación con la variable objetivo

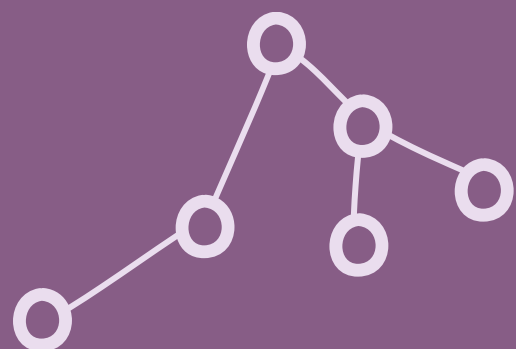


CART

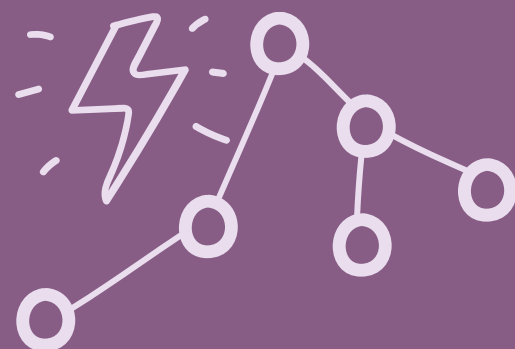
concave.points_worst	0.793566017
perimeter_worst	0.782914137
concave.points_mean	0.776613840
radius_worst	0.776453779
perimeter_mean	0.742635530
area_worst	0.733825035
radius_mean	0.730028511
area_mean	0.708983837
concavity_mean	0.696359707



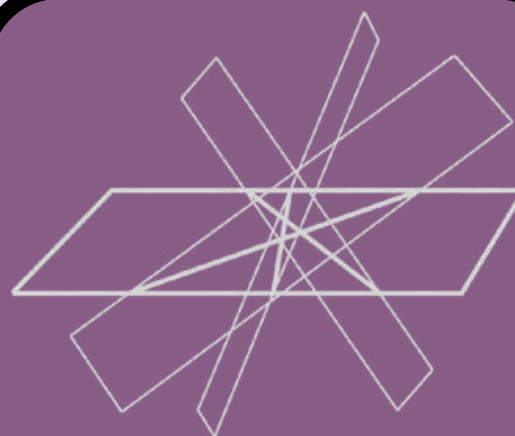
Modelos de aprendizaje supervisados



CART



**Random
Forest**



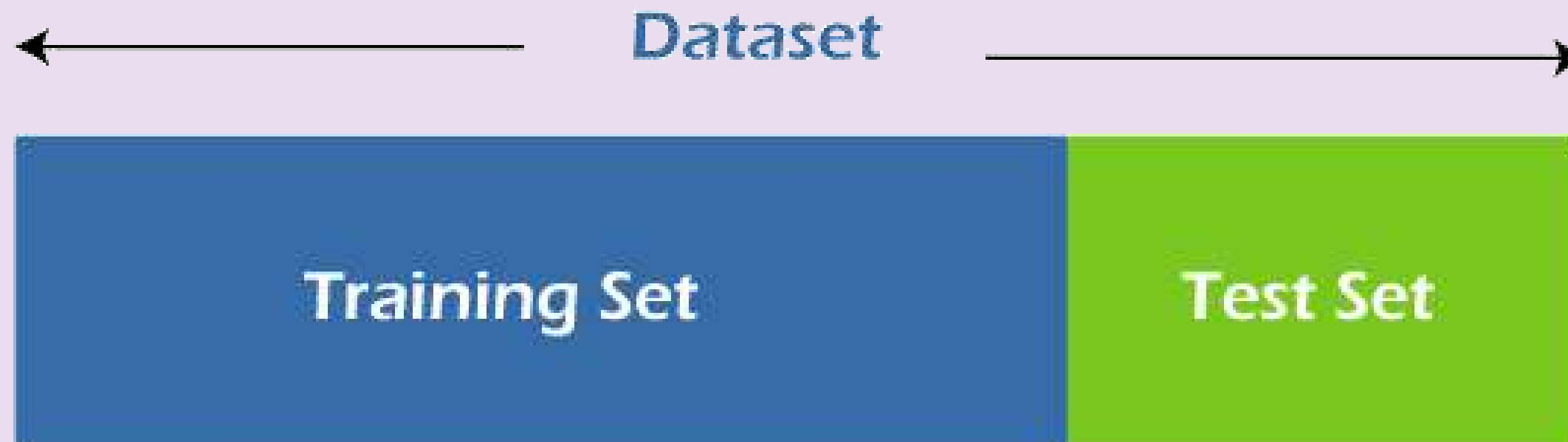
SVM



**Regresión
Logística**



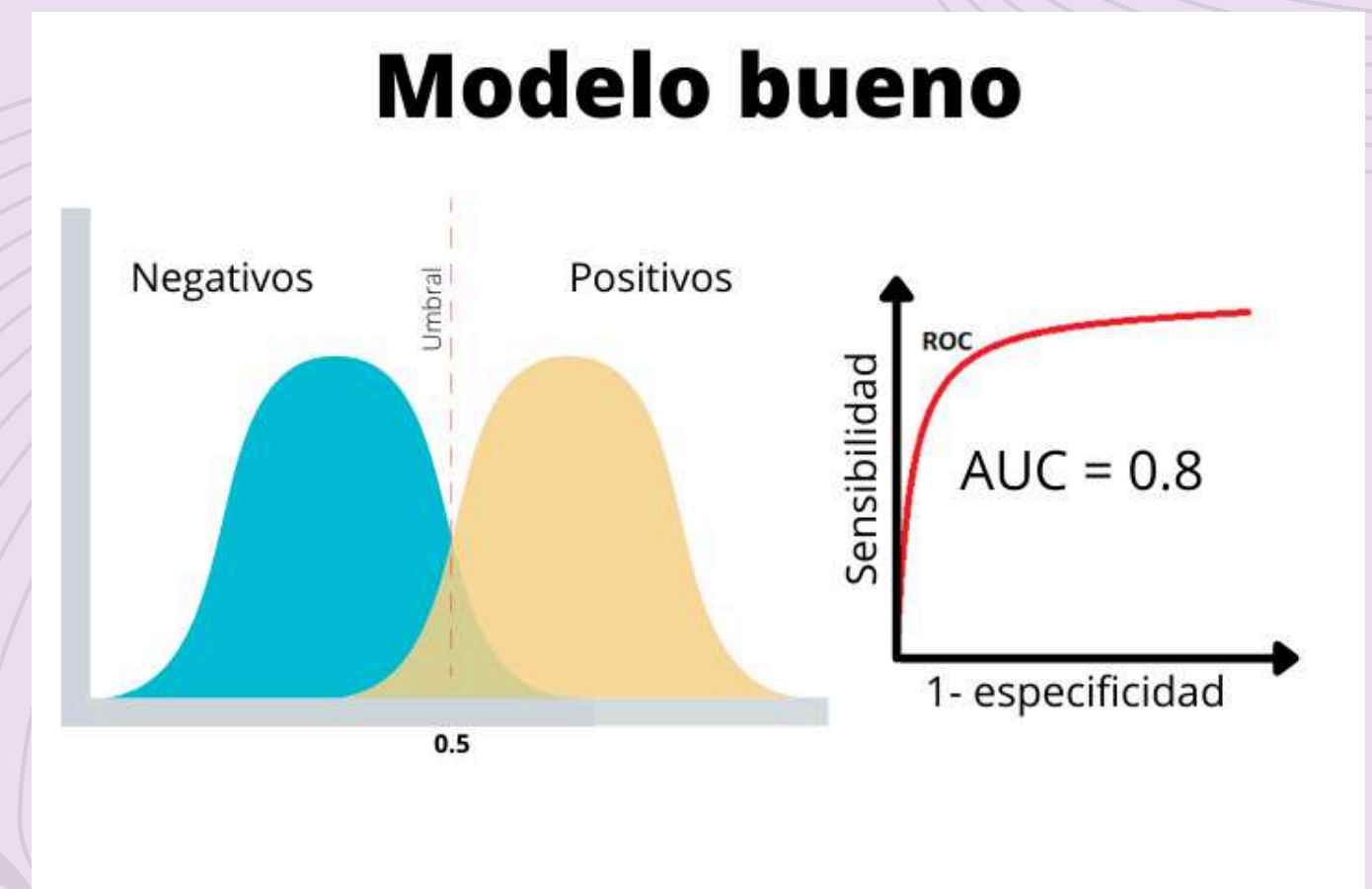
**Red
Neuronal**

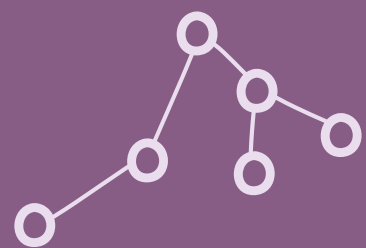


K-Fold con $K = 5$

Métricas de evaluación de rendimiento

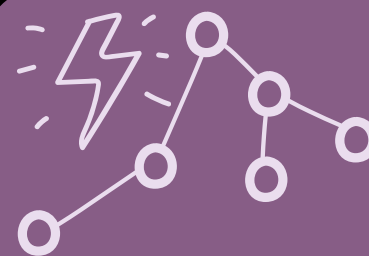
- ROC
- Sensibilidad
- Especificidad





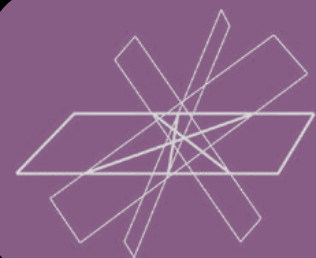
Árboles de Decisión (CART)

cp	ROC	Sens	Spec
0.004716981	0.9361931	0.9383412	0.8960133



Random Forest

mtry	ROC	Sens	Spec
2	0.9907537	0.9804382	0.9241417



Support Vector Machine (SVM)



C	ROC	Sens	Spec
0.25	0.9913307	0.9579812	0.9434109



Regresión Logística

ROC	Sens	Spec
0.9551611	0.9438185	0.9483942

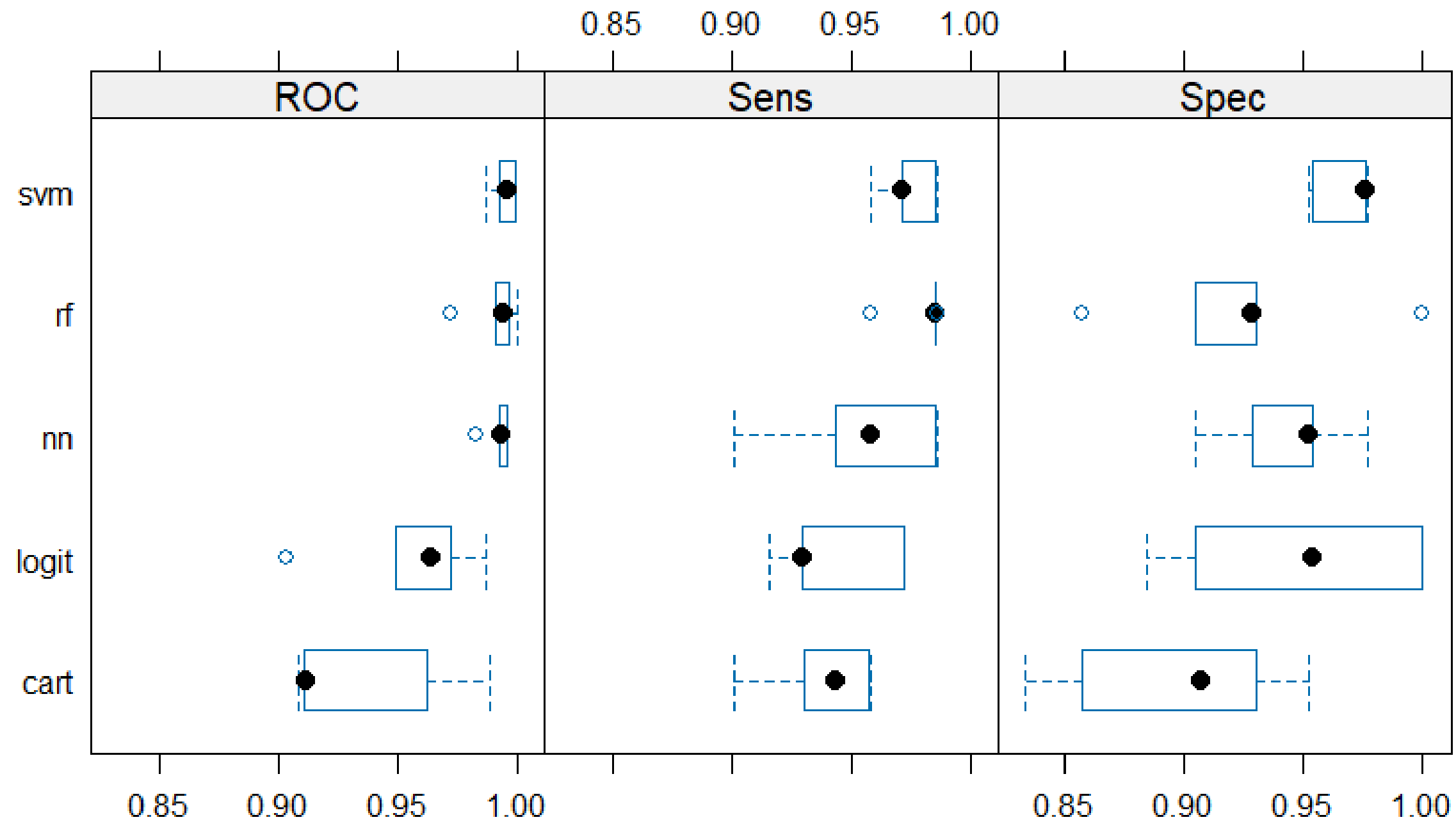


Redes Neuronales

ROC <dbl>	Sens <dbl>	Spec <dbl>	ROCSD <dbl>	SensSD <dbl>	SpecSD <dbl>
0.9867468	0.9719092	0.9341085	0.01192634	0.03153781	0.02541341



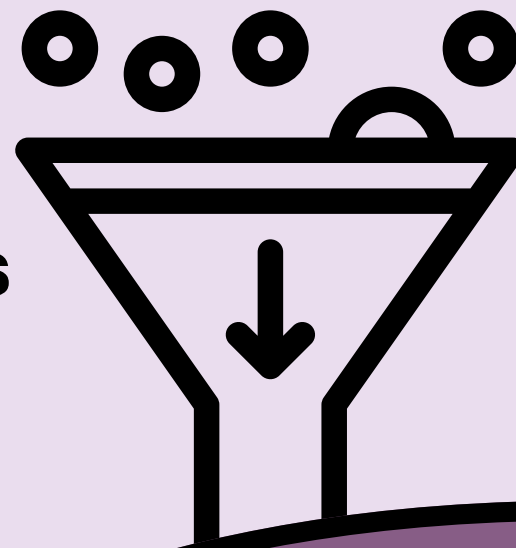
Comparación del rendimiento de los modelos





Importancia de variables en los distintos modelos

**Búsqueda de variables
más importantes**

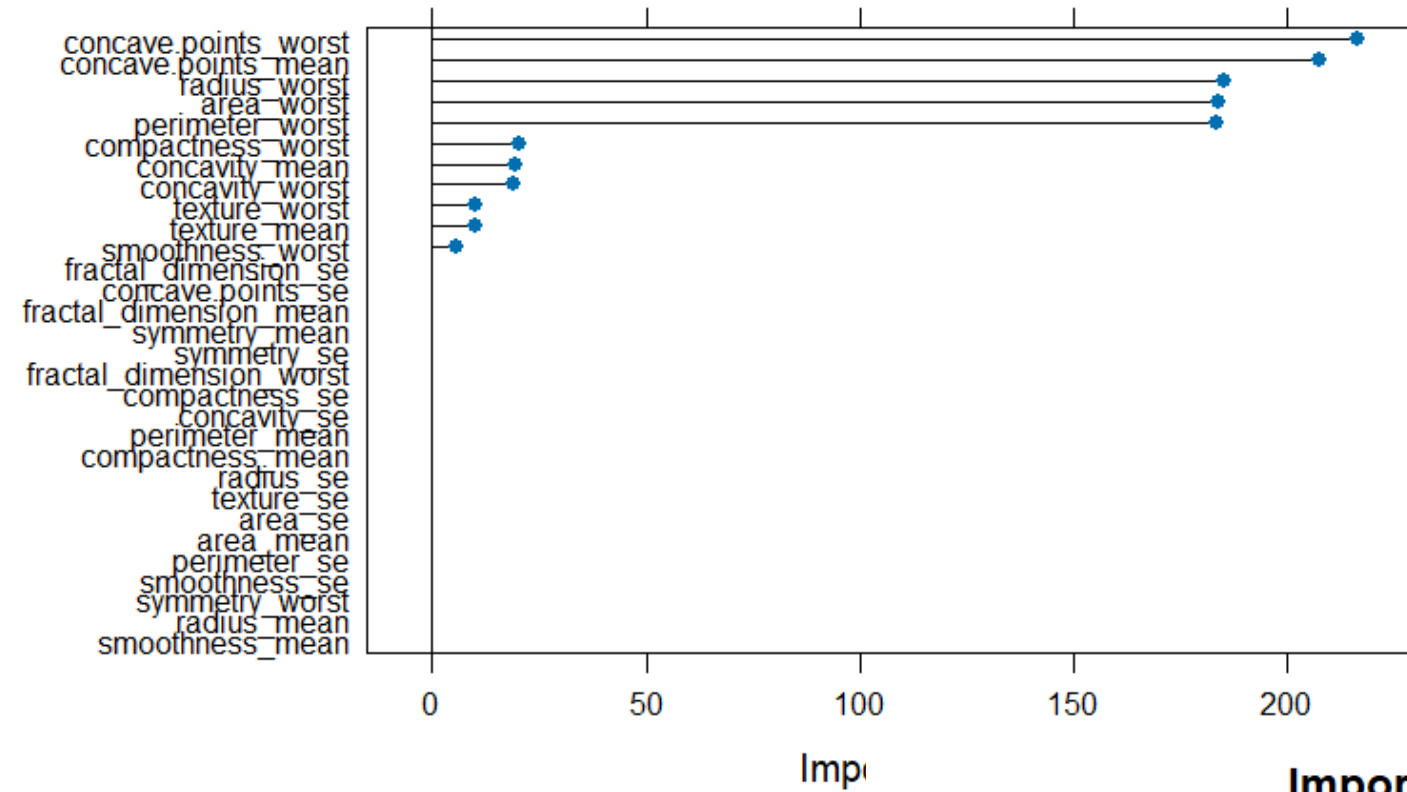




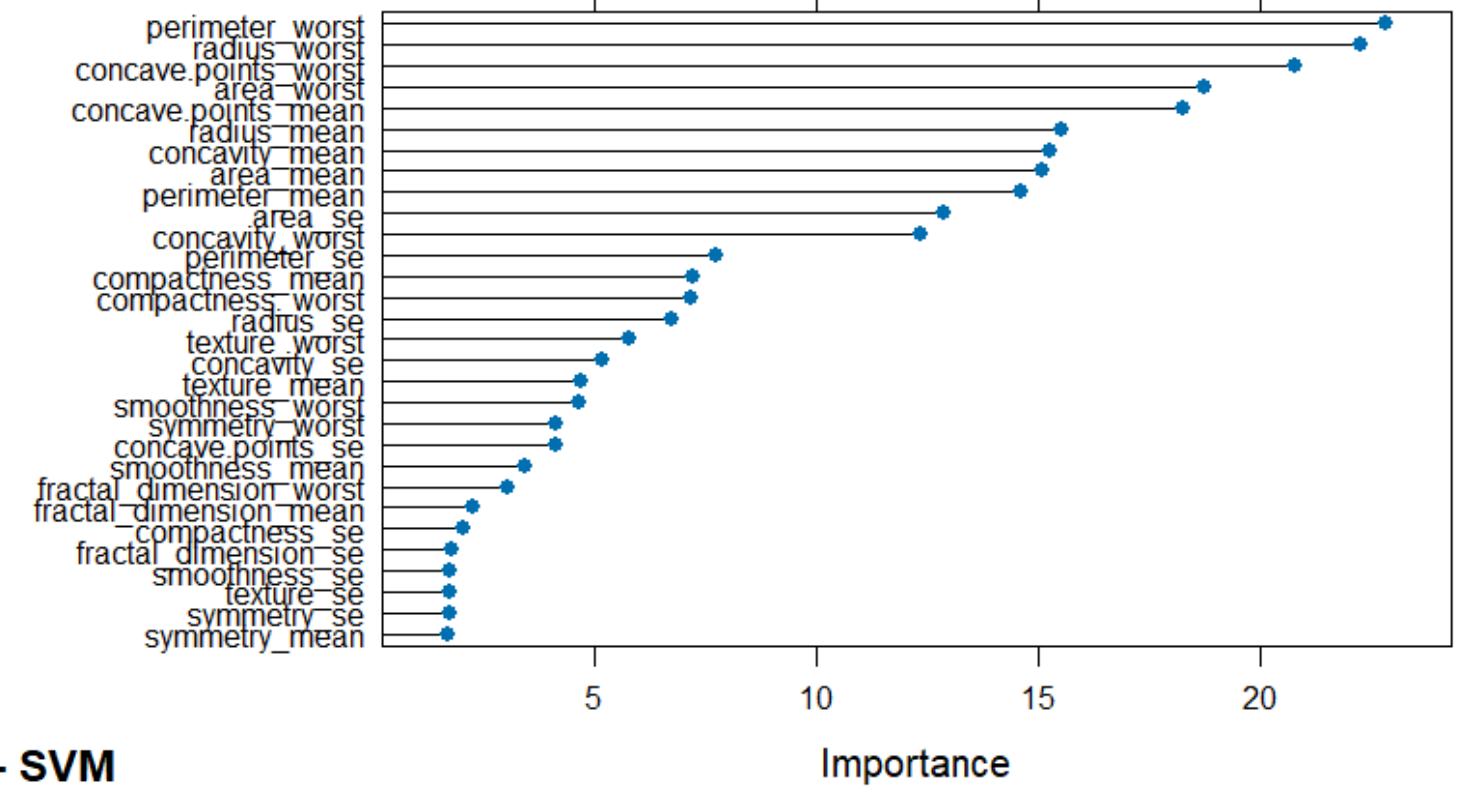
Modelos de aprendizaje supervisado con reducción de variables

5

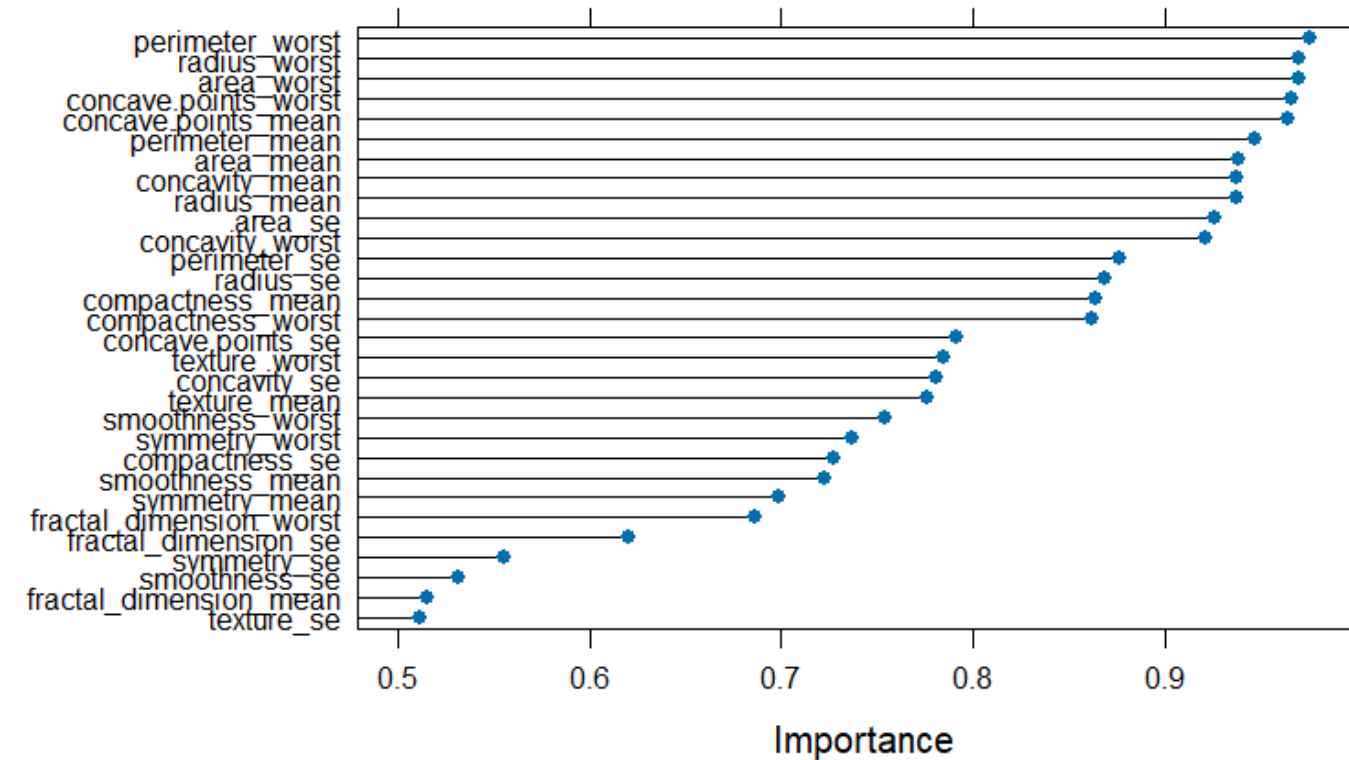
Importancia de Variables - CART



Importancia de Variables - Random Forest

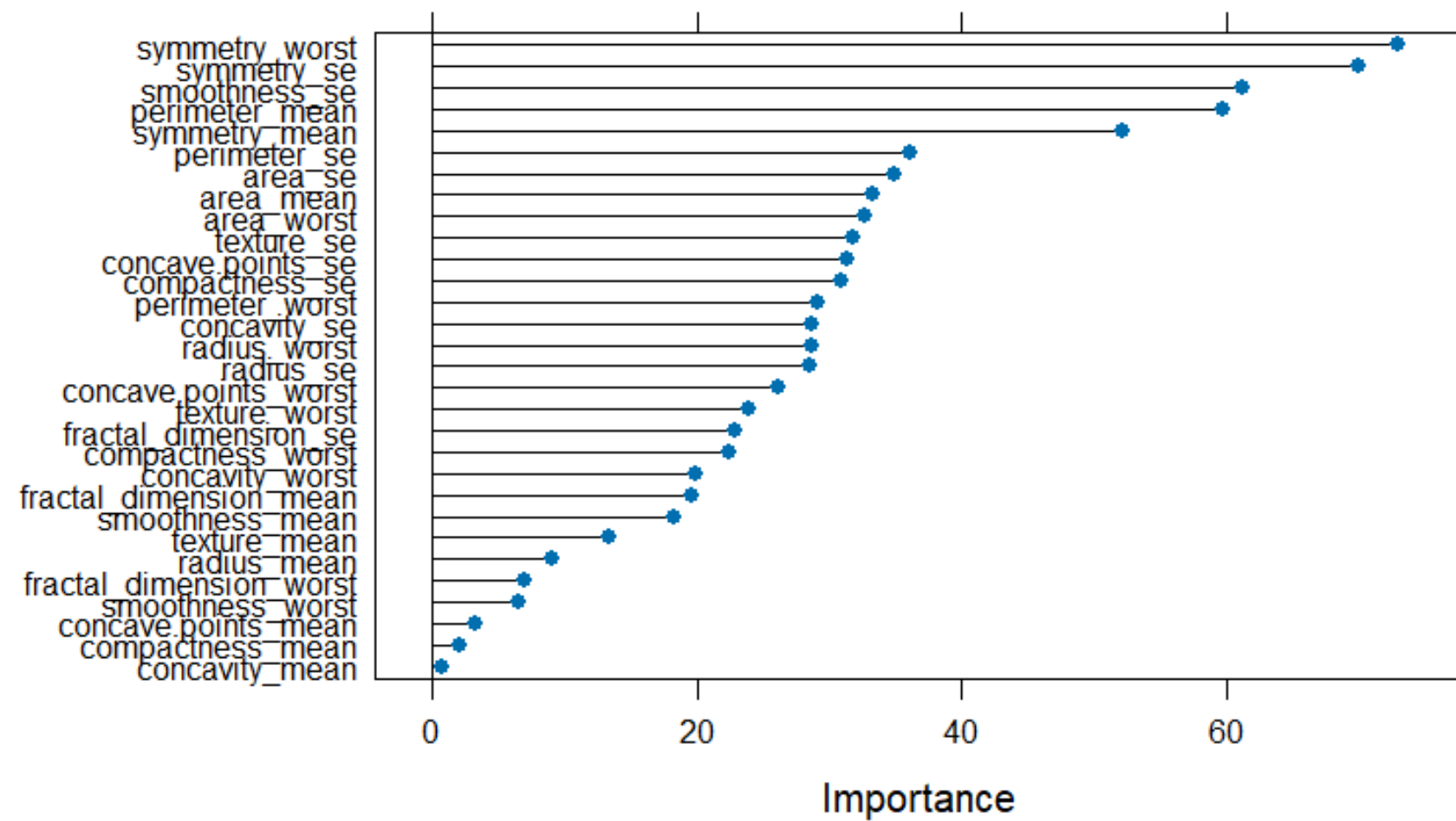


Importancia de Variables- SVM

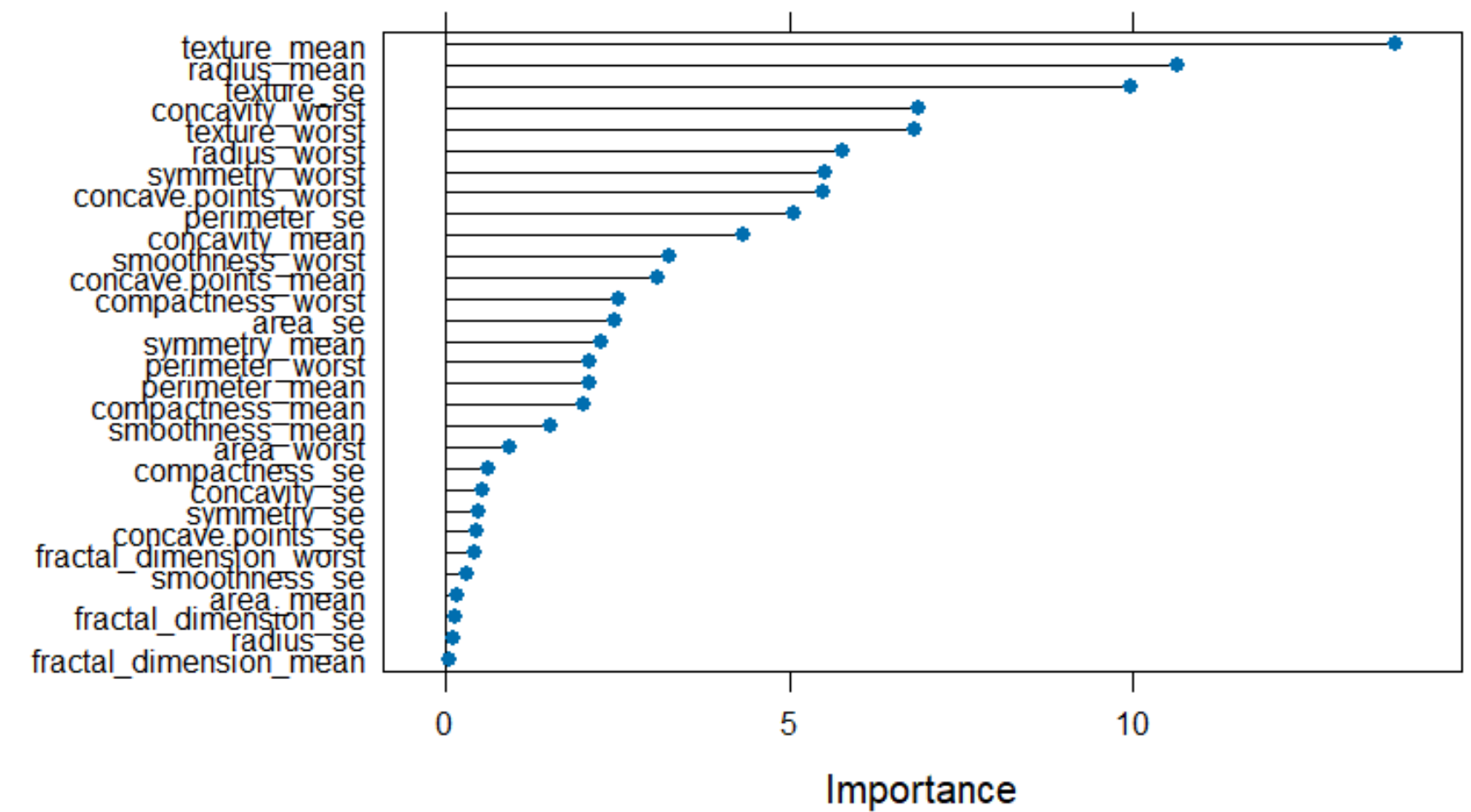




Importancia de Variables - GLM



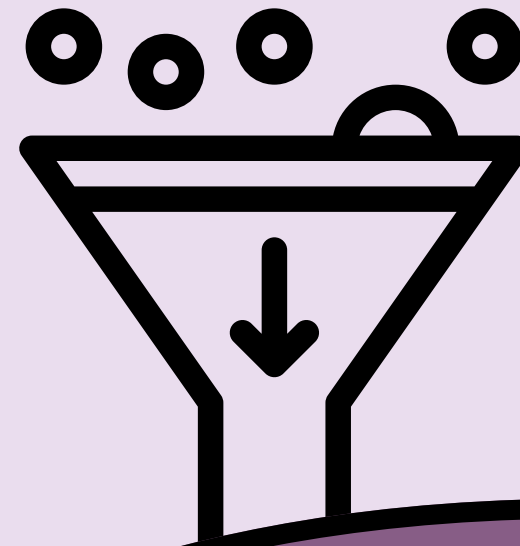
Importancia de Variables - Red Neuronal





Importancia de variables en los distintos modelos

**Búsqueda de variables
por encima del 90% de
importancia**



perimeter_worst

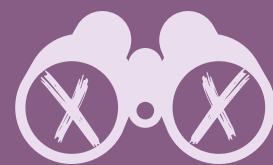
area_worst

concave.points_worst

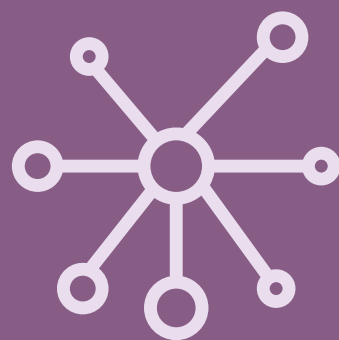


Comparación de resultados análisis supervisado dataset inicial vs dataset reducido

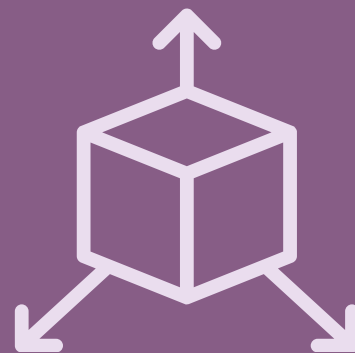
modelo	roc_change	sens_change	spec_change
cart	-0.0022295912	-0.014634757	-0.005067359
rf	0.001784723	0.0000000000	0.021090473
svm	0.001283809	0.002809440	-0.004924416
logit	0.012329249	0.006342232	0.004670715
nn	-0.003054248	0.003113223	-0.009862628



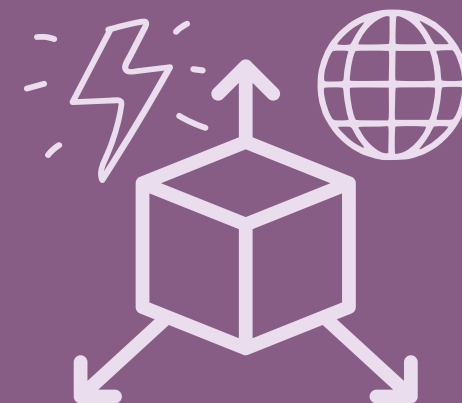
Modelos de aprendizaje no supervisados



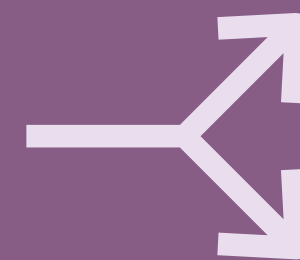
K-means
Jerarquico
DBSCAN



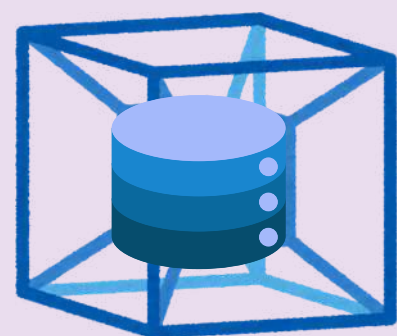
PCA



T-SNE



Reglas de
Asociación



31 - 1 columnas

569 filas

30 variables numéricas

~~**1 variable categórica**~~

~~**VALOR OBJETIVO**~~

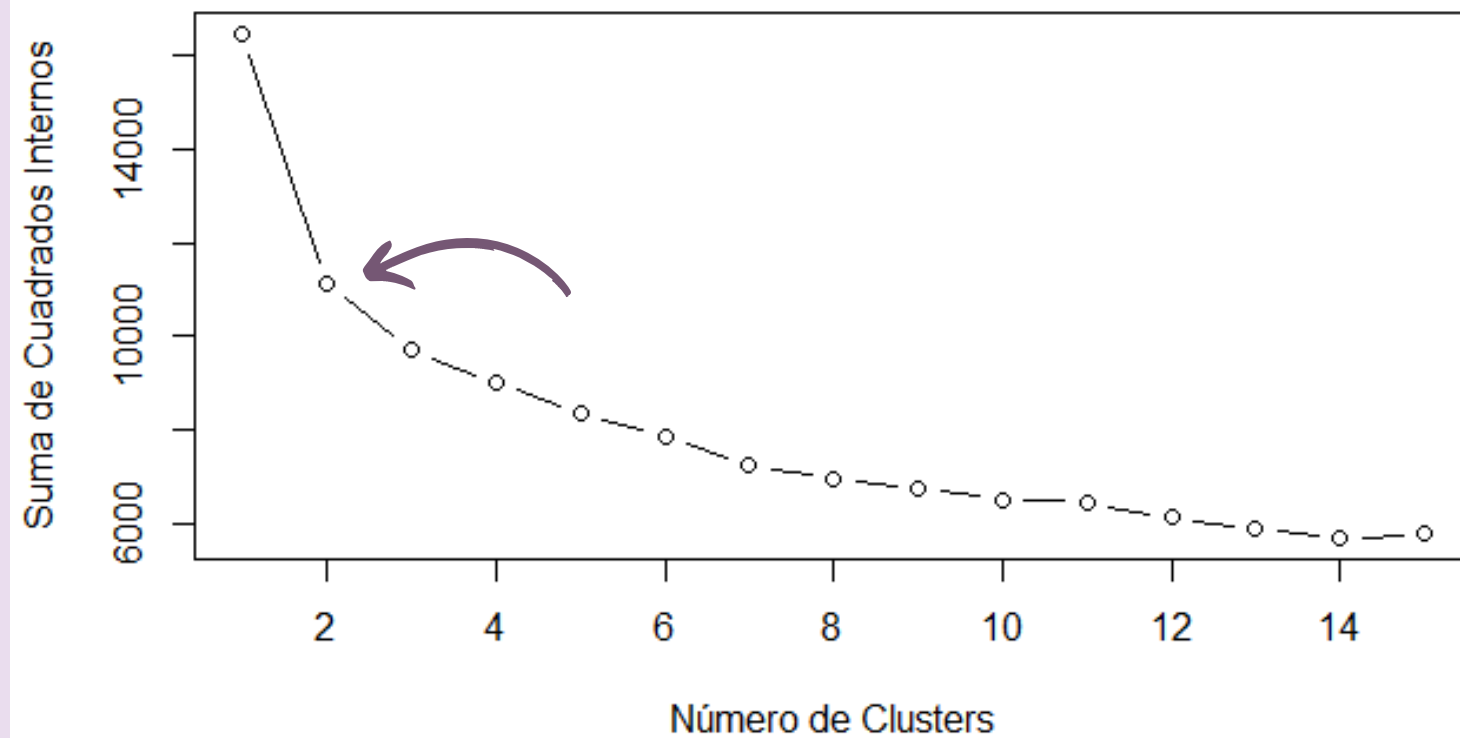


Normalización de los datos

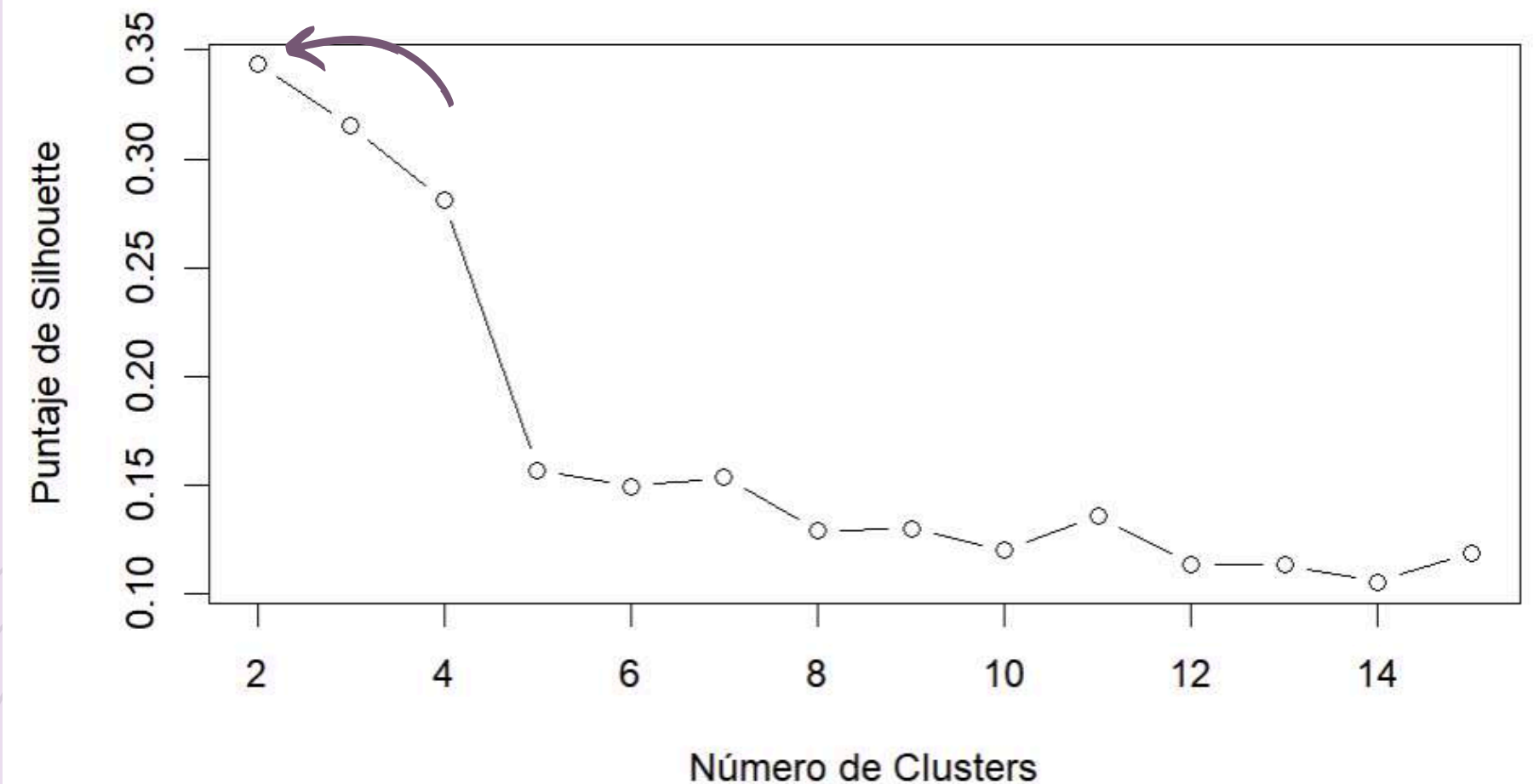


Clustering: Determinación del Número Óptimo de Clusters

Método del codo



Índice de Silhouette



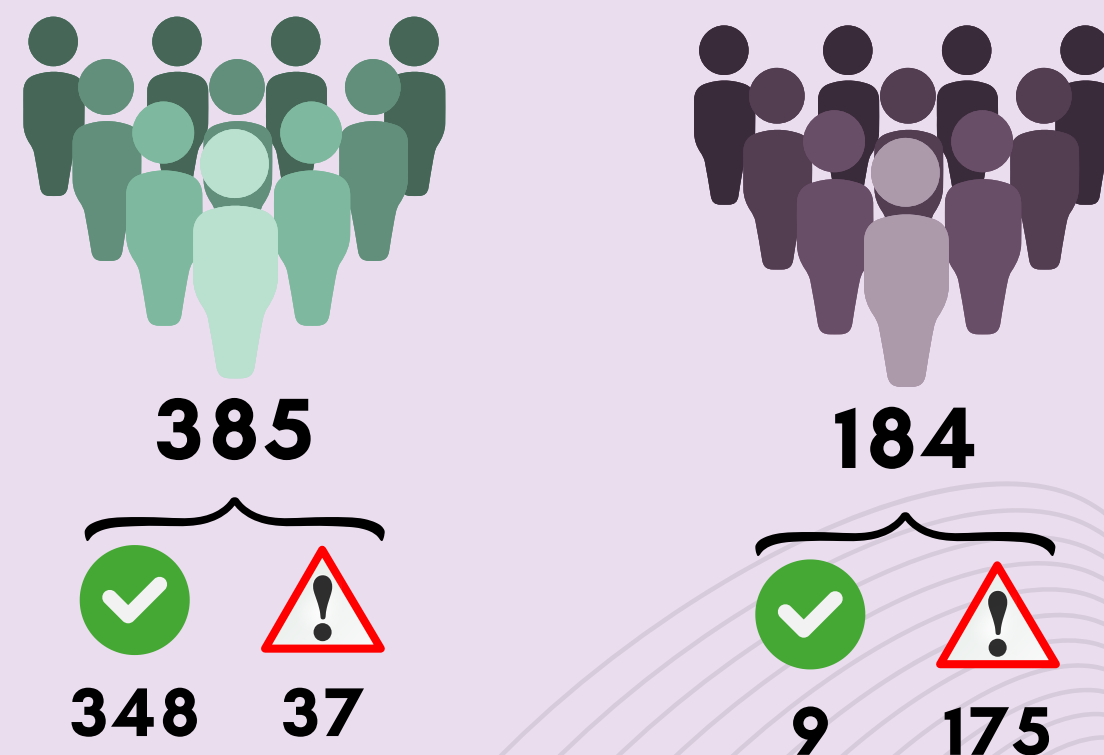
Número óptimo de clusters = 2



Algoritmo	Ancho de Silueta Promedio	Cohesión y Separación de Clusters
K-means	0.35	Alta cohesión y separación efectiva entre clusters
Clustering Jerárquico	0.29	Menor cohesión y separación comparado con K-means y DBSCAN
DBSCAN	0.35	Cohesión y separación similar a K-means, más efectiva que el jerárquico



K-Means con K = 2: organiza los datos en 2 clusters

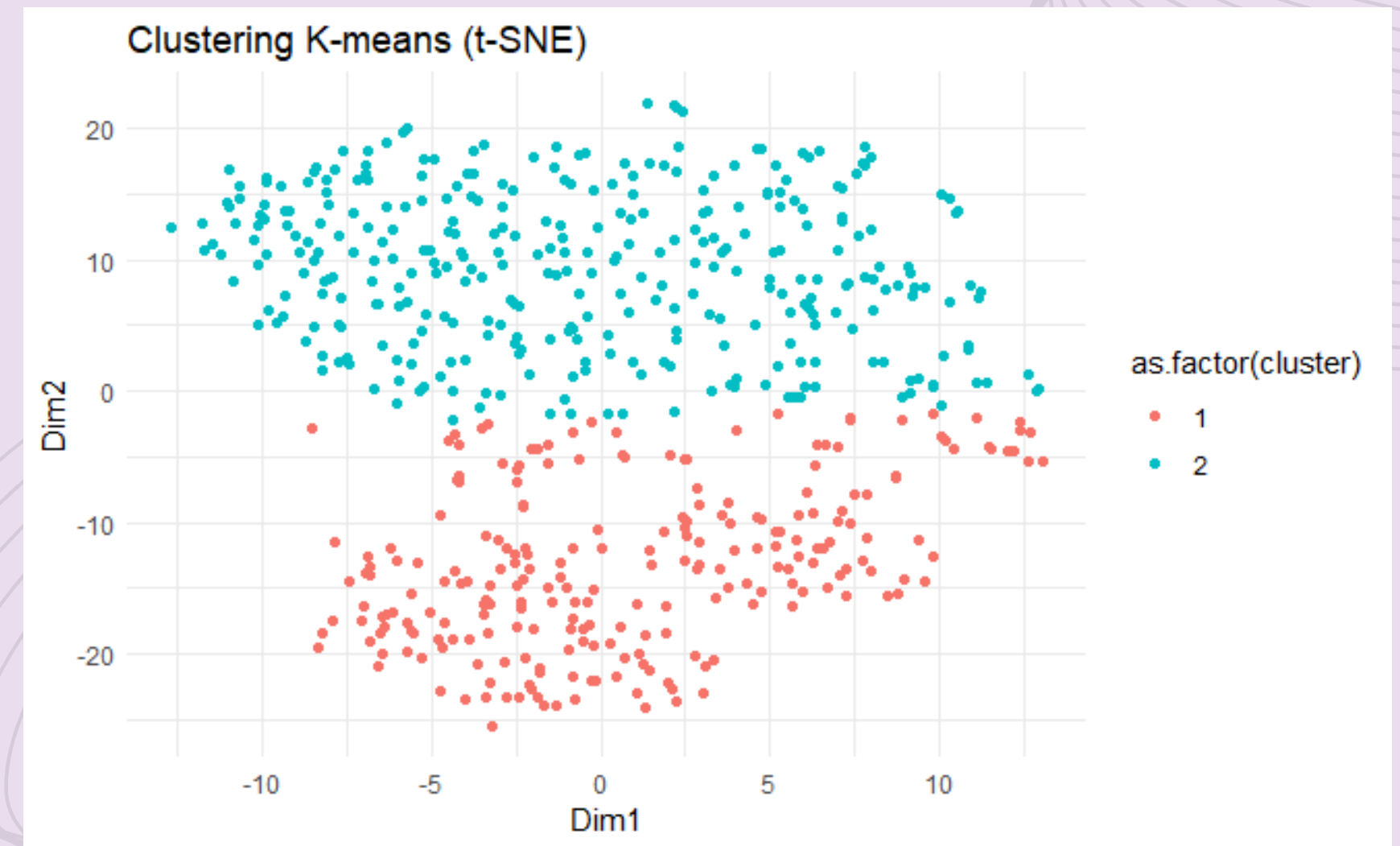
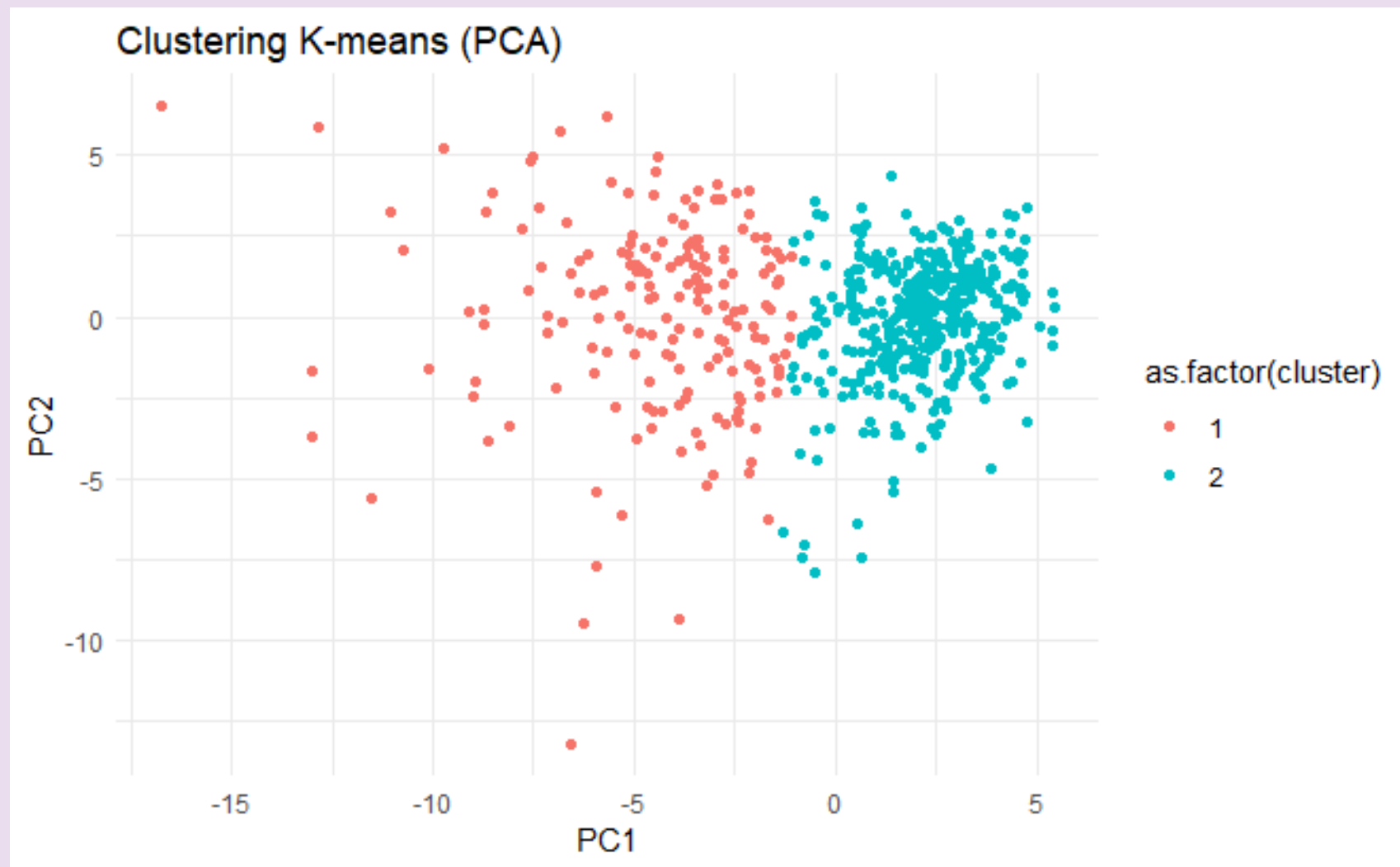


cluster <int>	diagnosis <dbl>	radius_mean <dbl>	texture_mean <dbl>	perimeter_ <dbl>	area_mean <dbl>
1	NA	12.39566	18.23860	79.67514	484.0797
2	NA	17.75054	21.48886	117.69266	1012.2891

2 rows | 1-6 of 32 columns



PCA Y T-SNE: Reducción de dimensinalidad






Comparación de resultados: Índice Silhouette

```
silhouette score - sin reducción: 0.3482031  
silhouette score - PCA: 0.3678096  
silhouette score - t-SNE: 0.5216206
```

Reglas de asociación con algoritmo Apriori



CONCLUSIONES



Conclusiones

**Modelos Random
Forest y SVM:**

**Rendimiento
excepcional**

`perimeter_worst`
`area_worst`
`concave.points_worst.`

No supervisado:
clustering con
reducción de
dimensionalidad
con t-SNE

**APLICACIÓN DE
DIVERSAS
TÉCNICAS DE
ANÁLISIS**