

# Clasificación de Parkinson mediante **voz**



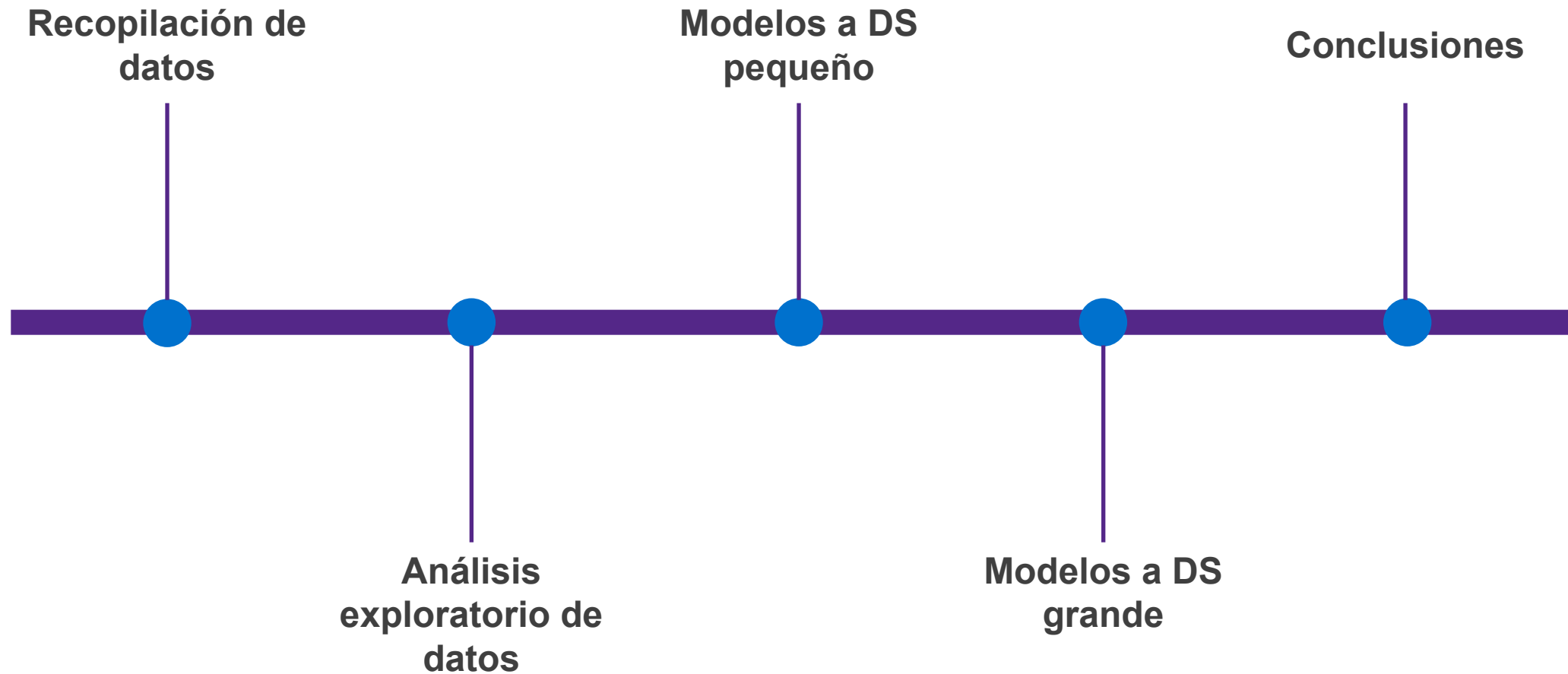
UNIVERSITAT DE  
BARCELONA

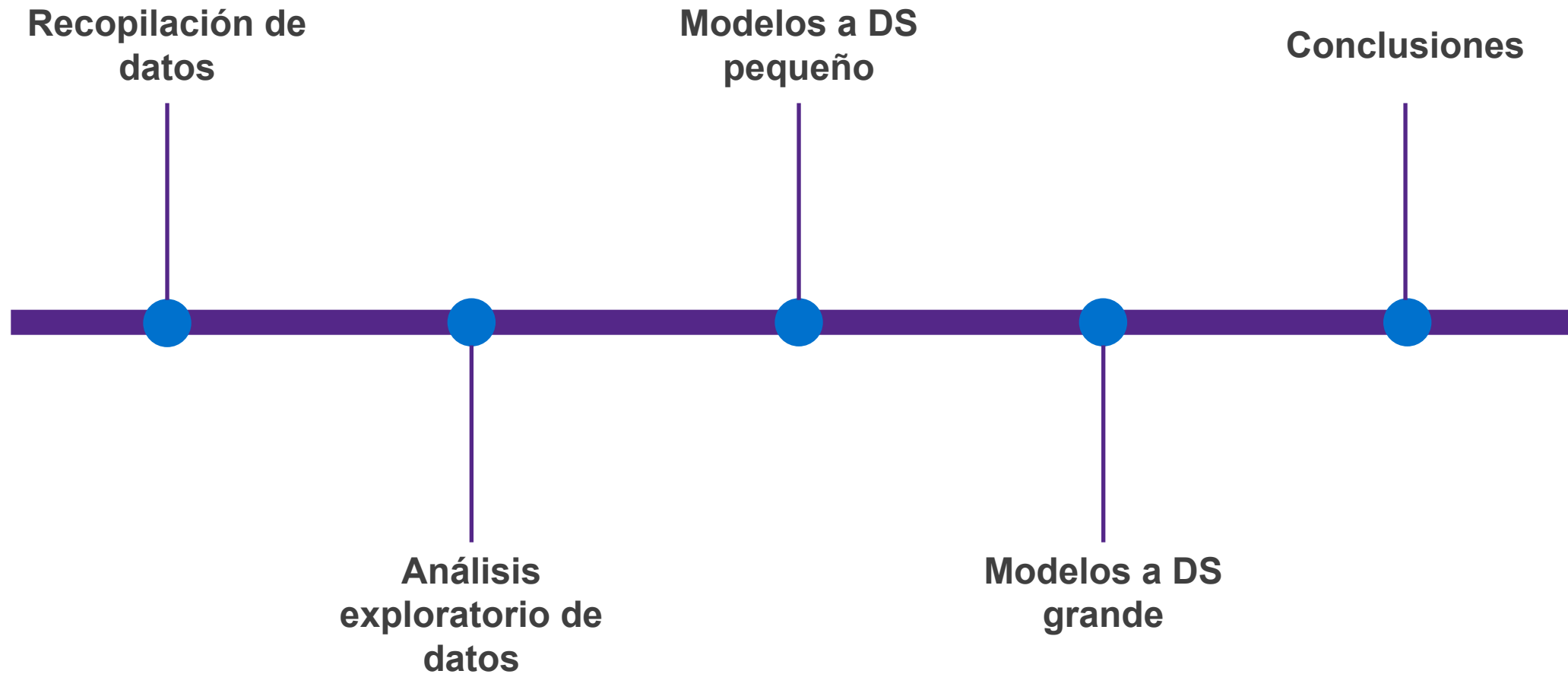


DATA SCIENCE @ UNIVERSITAT DE BARCELONA

### ¿Se puede predecir si un paciente tiene Parkinson mediante el análisis de grabaciones de voz?

- Búsqueda de datasets que tengan grabaciones de voz de personas con Parkinson
- Análisis de las variables aportadas para comprobar cuáles son las que tienen más peso en su clasificación
- Aplicación de métodos de clasificación para diagnosticar personas con Parkinson.
- Dar respuesta a nuestra pregunta: ¿es posible?





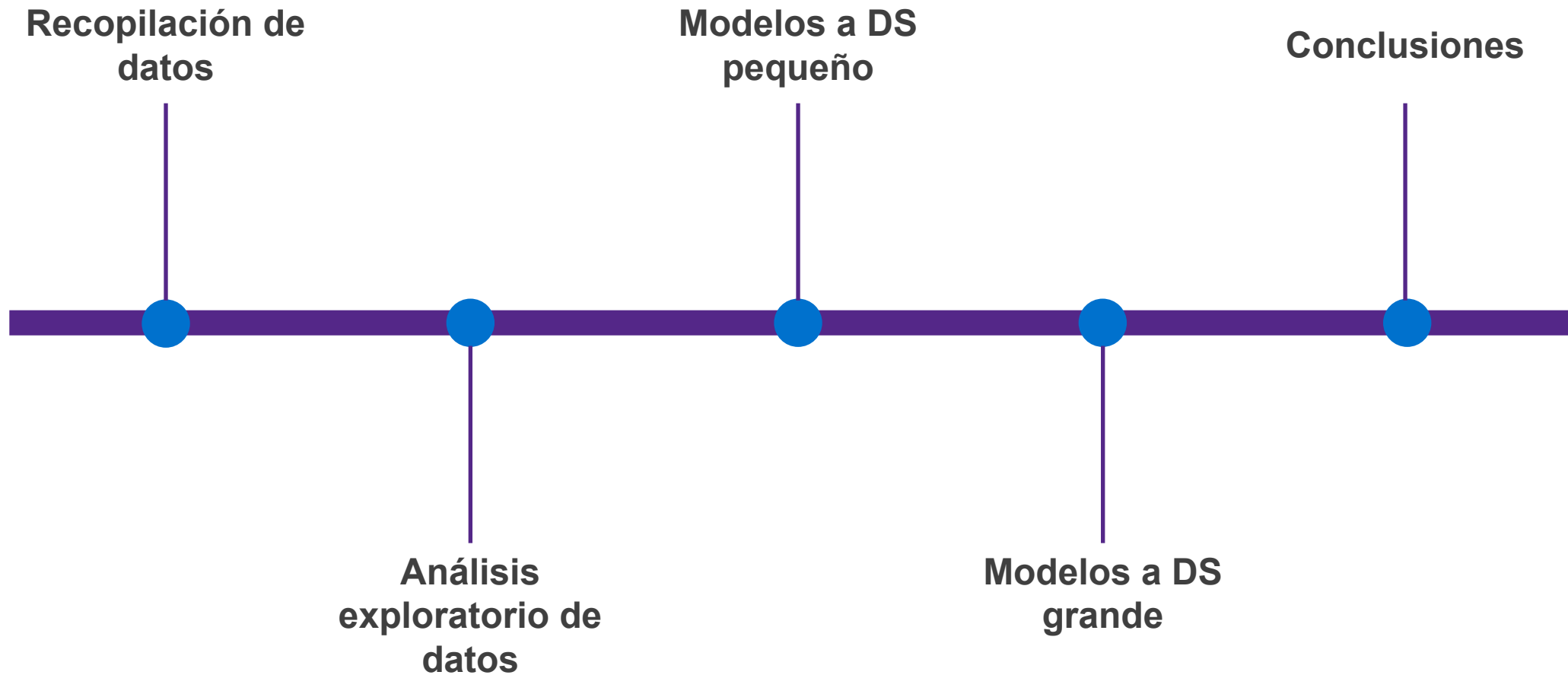
### Búsqueda de datasets sobre Parkinson

#### Basado en medidas de voz (Pequeño)

- Encontrado en Machine Learning Repository
- 31 pacientes con 195 grabaciones de audio
- Datos preprocesados

#### Basado en medidas de voz (Grande)

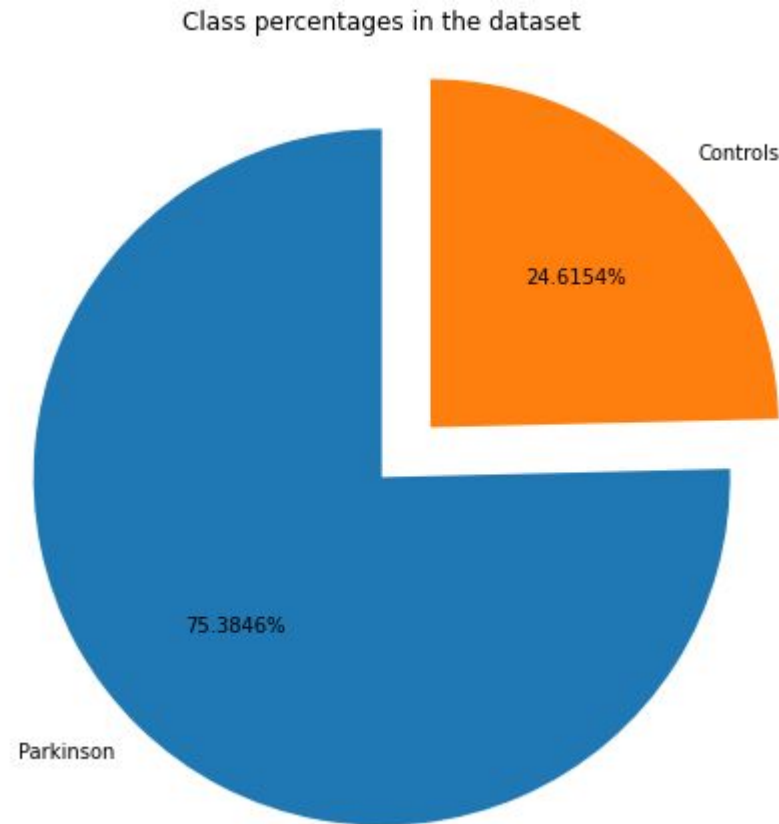
- Encontrado en Machine Learning Repository
- 252 pacientes con 755 grabaciones de audio
- Datos preprocesados



### Analizamos el dataset de voz (pequeño)

- **name** - código del sujeto y de la grabación
- **MDVP:Fo(Hz)** - Frecuencia fundamental vocal media
- **MDVP:Fhi(Hz)** - Frecuencia fundamental vocal máxima
- **MDVP:Flo(Hz)** - Frecuencia fundamental vocal mínima
- **MDVP:Jitter(%)**, **MDVP:Jitter(Abs)**, **MDVP:RAP**, **MDVP:PPQ**, **Jitter:DDP** - Diversas medidas que miden la variación en la frecuencia fundamental
- **MDVP:Shimmer**, **MDVP:Shimmer(dB)**, **Shimmer:APQ3**, **Shimmer:APQ5**, **MDVP:APQ**, **Shimmer:DDA** - Diversas medidas que miden variación en la amplitud
- **NHR**, **HNR** - Dos medidas que miden el ratio ruido-armónicos (Noise to Harmonic Ratio, NHR) y el ratio armónicos-ruido (Harmonic to Noise Ratio, HNR)
- **status** - Indica el estado de salud del sujeto: sano (0) o con Parkinson (1)
- **RPDE**, **D2** - Dos medidas de complejidad dinámica no lineales
- **DFA** - Exponente de escala fractal de señal
- **spread1**, **spread2**, **PPE** - Tres medidas no lineales de variación en frecuencia fundamental

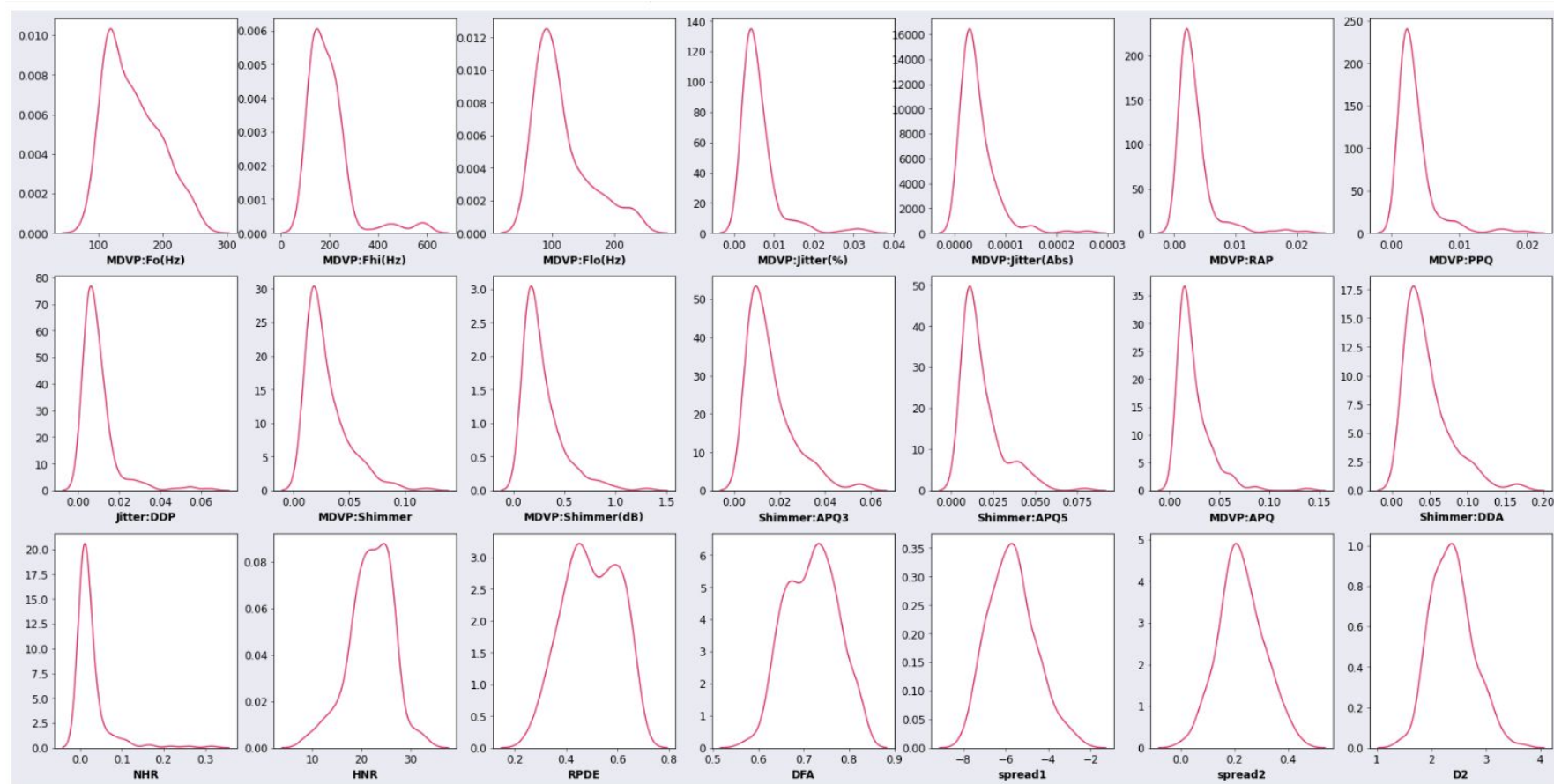
### Analizamos el dataset de voz (pequeño) – Distribución del dataset



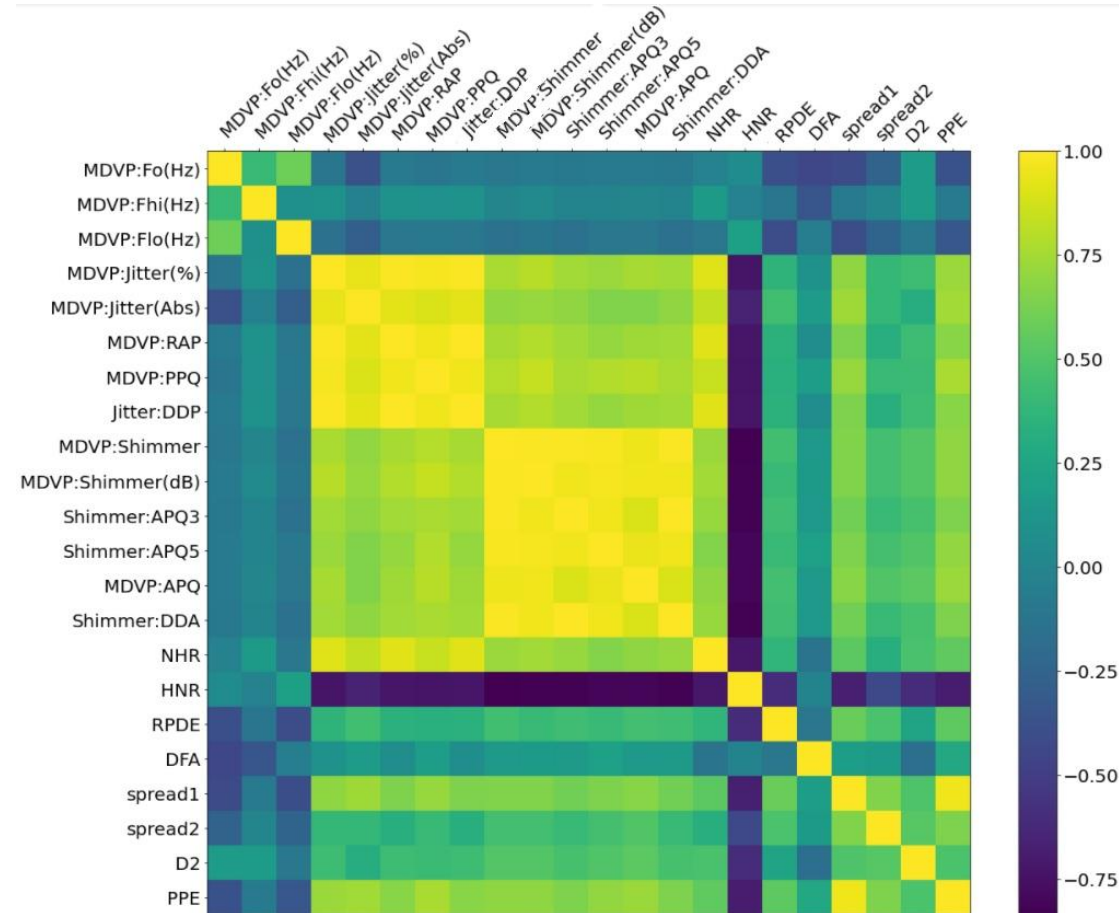


### Analizamos el dataset de voz (pequeño) – Distribución de las variables

- La mayoría de las variables tienen una distribución normal



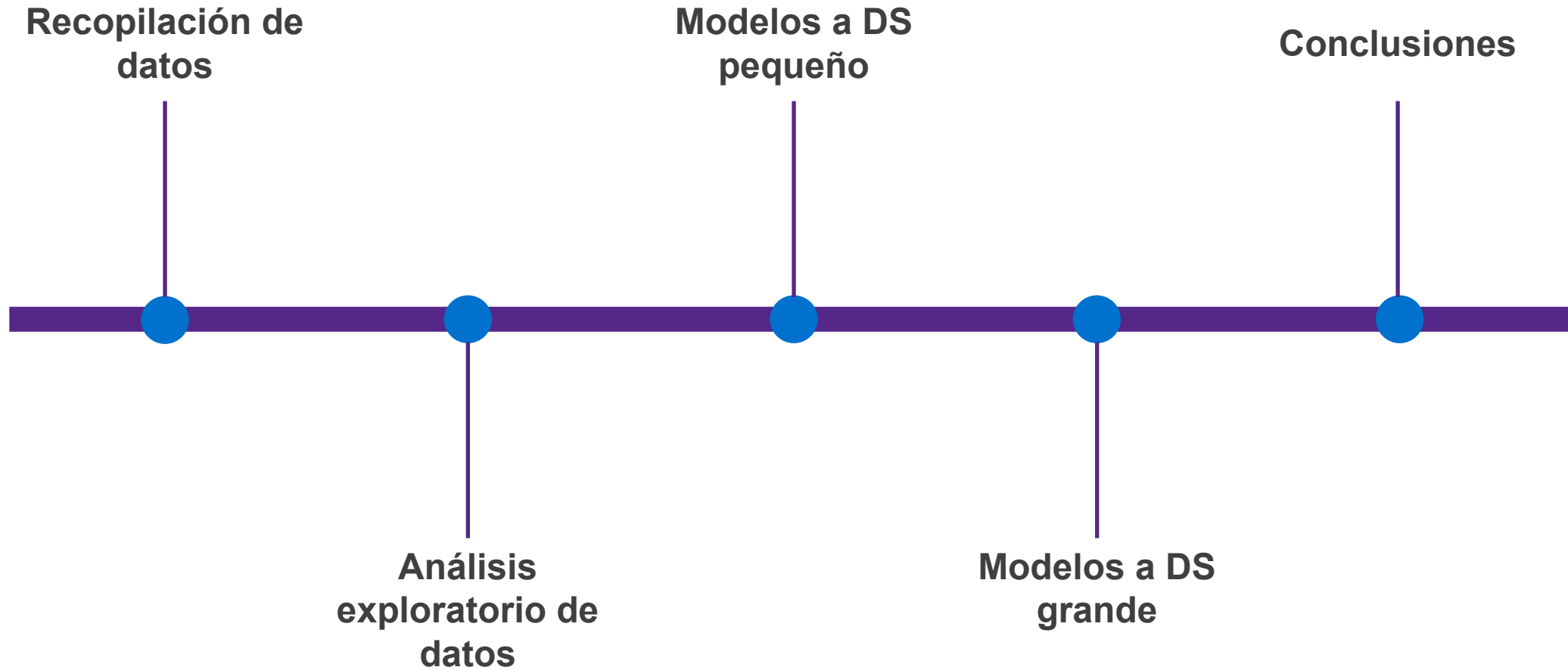
### Analizamos el dataset de voz (pequeño) – Correlación de las variables



### Analizamos el dataset de voz (pequeño) – Conclusiones

- La mayoría de variables presentación distribución normal
- Todas las medidas de “*Several measures of variation in fundamental frequency*” (MDVP:Jitter(%), MDVP:Jitter(Abs), MDVP:RAP, MDVP:PPQ, Jitter:DDP) y “*Several measures of variation in amplitude*” (MDVP:Shimmer, MDVP:Shimmer(dB), Shimmer:APQ3, Shimmer:APQ5, MDVP:APQ, Shimmer:DDA) están altamente y positivamente correlacionadas
- Las variables HNR, RFDE, spread1 y spread2 parecen ser las que más afecten – presentan diferencias entre personas control y personas con Parkinson
- En definitiva, muchas variables están correlacionadas – Posibilidad aplicar medidas para evitar colinealidad

**Finalmente no quitamos variables por la poca cantidad de datos en el dataset**

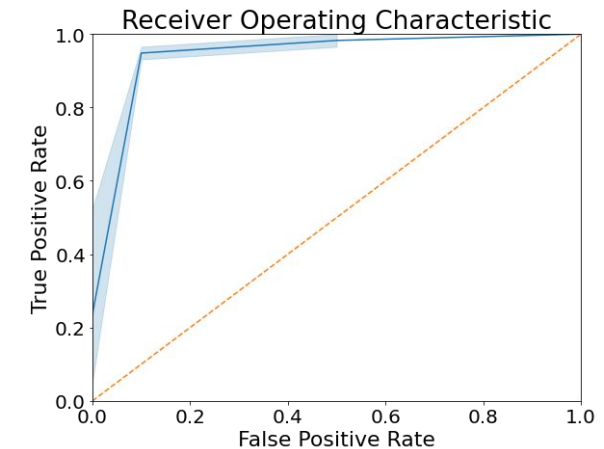
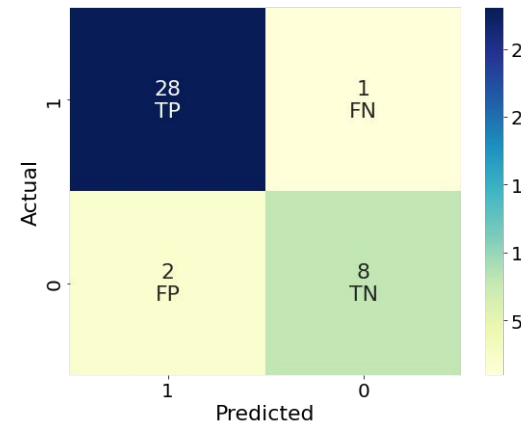


### Comprobamos varios modelos con el dataset de voz (pequeño)

#### Accuracies:

- **KNN-3** = 0.8974358974358975
- **SVM** = 0.8461538461538461
- **NB** = 0.7948717948717948
- **RF** = 0.9230769230769231
- **BC** = 0.8717948717948718
- **XGB** = 0.8461538461538461
- **CAT** = 0.9230769230769231

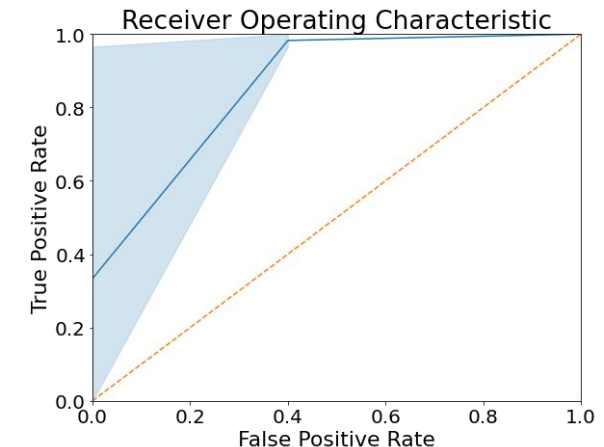
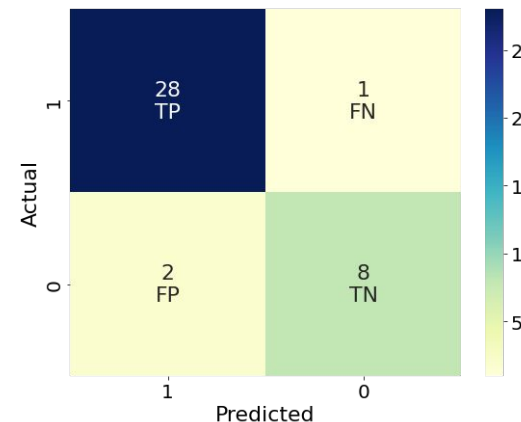
#### Random Forest (RF)



#### Area under the curve:

- **KNN-3** = 0.8327586206896551
- **SVM** = 0.7327586206896552
- **NB** = 0.8620689655172413
- **RF** = 0.8827586206896553
- **BC** = 0.7827586206896553
- **XGB** = 0.7327586206896552
- **CAT** = 0.8827586206896553

#### Catboost (CAT)

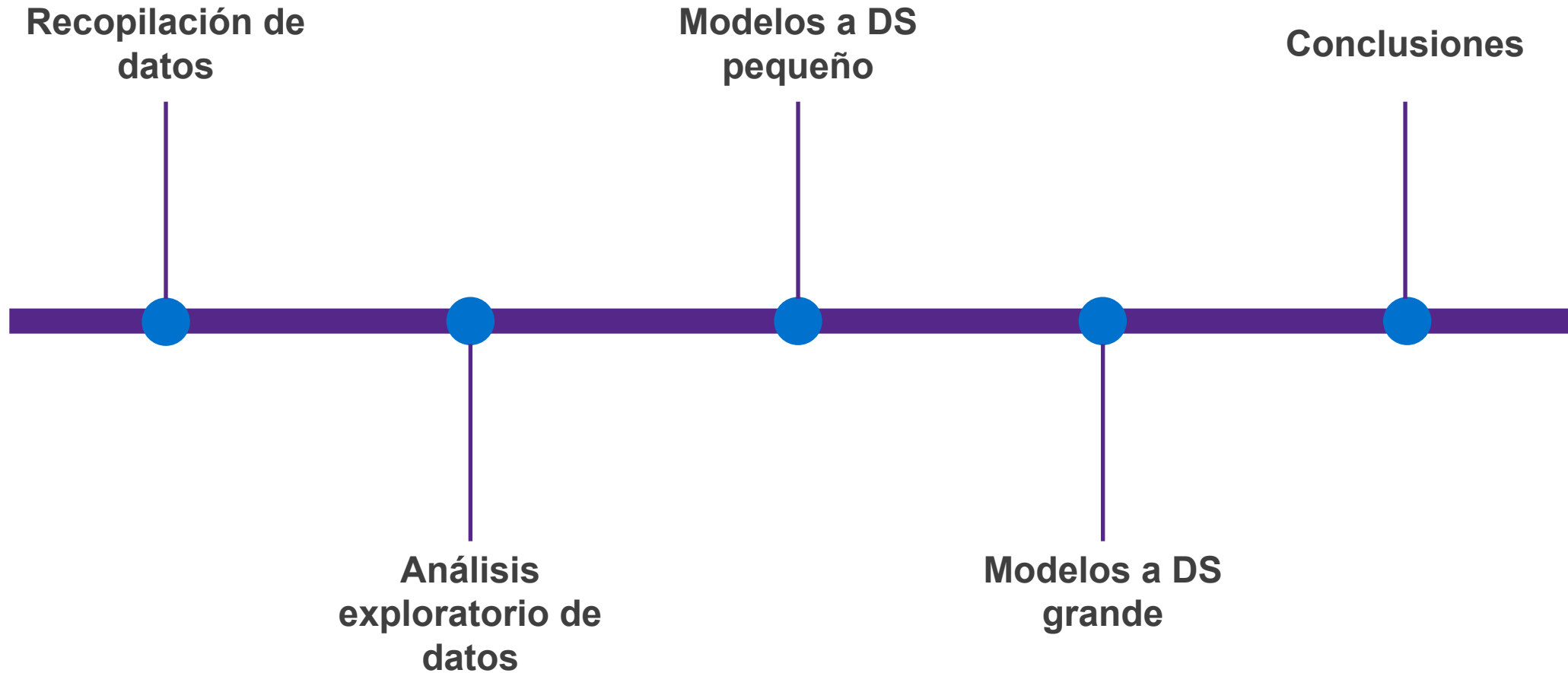


### Conclusiones

- Mejor modelos: **Random Forest y Catboost** (Accuracy: 0.92, Area under the curve: 0.88)
  - *Resultados iguales quizá por la similaridad entre los dos modelos (ambos consisten de un conjunto de decision trees).*
- Posibles mejoras del modelo:
  - Conseguir más datos
  - Reducir el número de variables – Descartado por dataset pequeño
  - Aplicar SMOTE – Igualar número de pacientes control y pacientes con Parkinson
  - Modificar la preparación del train y test, evitando que un mismo paciente con varias grabaciones se divida aleatoriamente entre los dos
  - Aplicar modelos de Deep Learning – Descartado por dataset pequeño

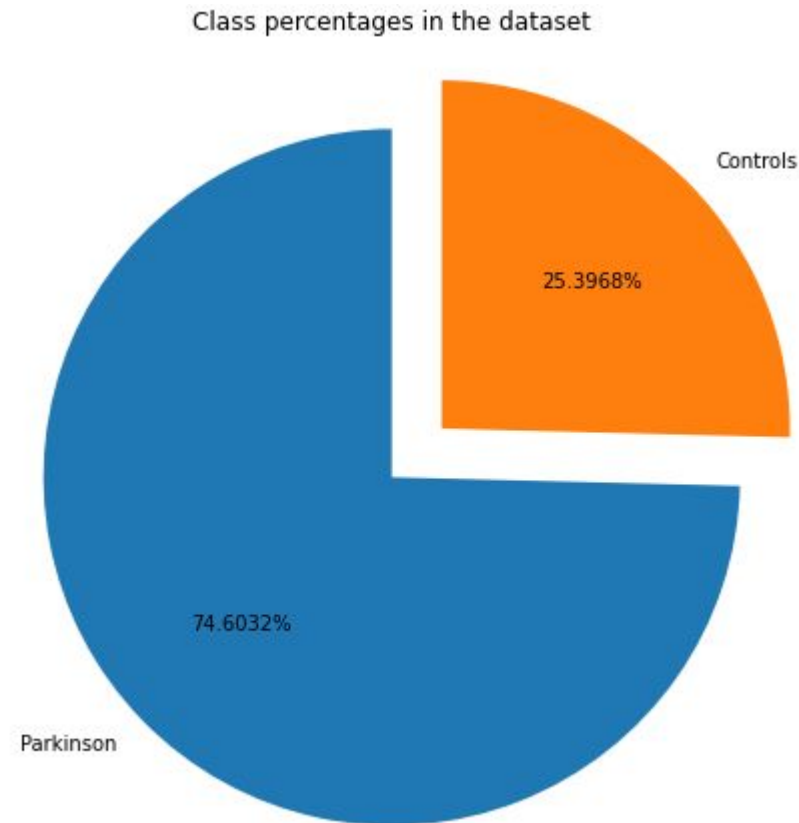
## Conclusiones

- Mejor modelos: **Random Forest y Catboost** (Accuracy: 0.92, Area under the curve: 0.88)
  - *Resultados iguales quizá por la similaridad entre los dos modelos (ambos consisten de un conjunto de decision trees). Catboost parece alcanzar siempre valores más altos de Accuracy y AUC ya que al crear un decision tree nuevo, toma en consideración los resultados del decision tree anterior, de manera serial. En cambio, RF avalua los diferentes decision trees de manera paralela.*
- Posibles mejoras del modelo:
  - Conseguir más datos
  - Reducir el número de variables – Descartado por dataset pequeño
  - **Aplicar SMOTE – Igualar número de pacientes control y pacientes con Parkinson**
  - **Modificar la preparación del train y test, evitando que un mismo paciente con varias grabaciones se divida aleatoriamente entre los dos**
  - Aplicar modelos de Deep Learning – Descartado por dataset pequeño









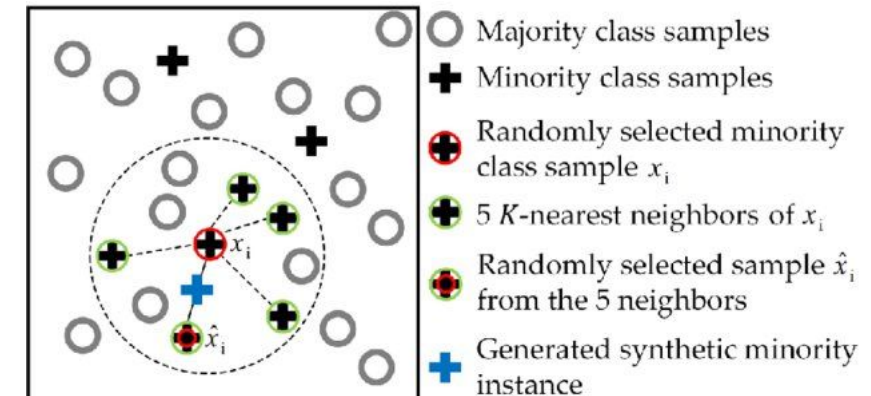
Con lo aprendido en anterior dataset pequeño, análisis del dataset de voz (grande)



### ¿Qué es SMOTE (*Synthetic Minority Over-sampling Technique*)?

- Se usa cuando - en un problema de clasificación - las clases que hay que discriminar no están representadas proporcionalmente. Si no tenemos en cuenta este desequilibrio, el clasificador tenderá a predecir la clase mayoritaria.
- SMOTE genera de forma sintética nuevos elementos de la clase minoritaria usando como referencia los elementos de dicha clase ya presentes en el conjunto de datos. ¿Cómo?

- Elige un elemento de la clase minoritaria al azar 
- Escoge un número de vecinos cercanos a este elemento 
- Elige uno de estos vecinos 
- Se genera un nuevo elemento de la clase minoritaria 



Ma et al., 2019, [10.3390/rs11070846](https://doi.org/10.3390/rs11070846)

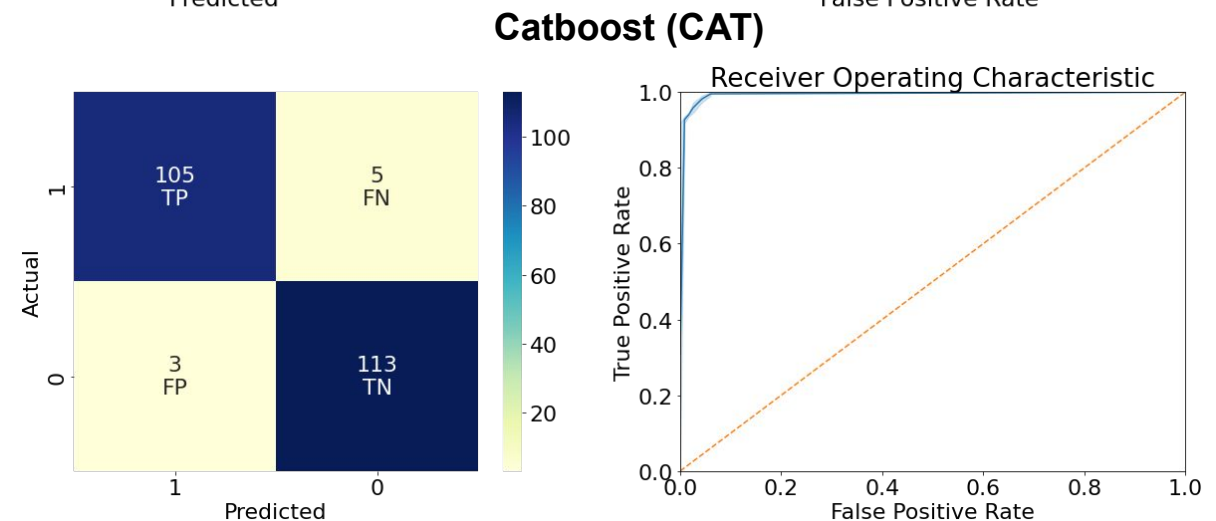
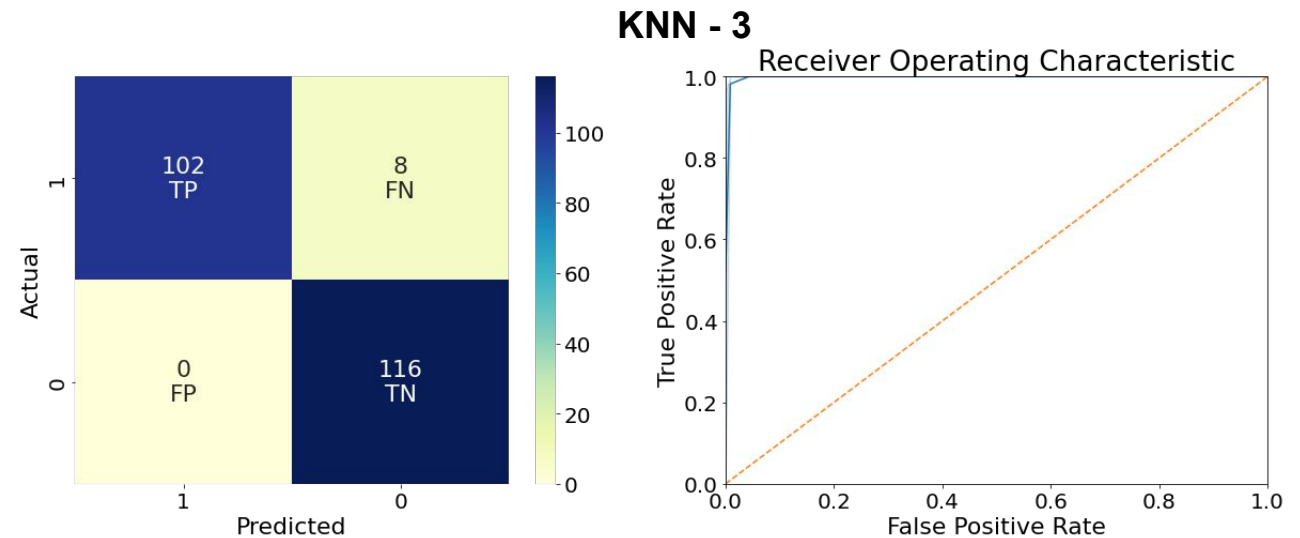
### Comprobamos varios modelos con el dataset de voz (grande) después de aplicar SMOTE

#### Accuracies (con SMOTE):

- KNN-3 = 0.9646017699115044
- SVM = 0.9336283185840708
- NB = 0.8141592920353983
- RF = 0.9336283185840708
- BC = 0.911504424778761
- XGB = 0.9513274336283186
- Catboost = 0.9646017699115044

#### Area under the curve (con SMOTE):

- KNN-3 = 0.9636363636363636
- SVM = 0.9325235109717868
- NB = 0.8140282131661443
- RF = 0.9325235109717868
- BC = 0.9100313479623825
- XGB = 0.9509404388714734
- Catboost = 0.9643416927899686



### ¿Qué es el método StratifiedGroupKFold?

Es una técnica de validación cruzada que realiza:

- Tratamiento de los datos en grupos (Sujetos) para mover un sujeto entero a training o test y forzar que todas las observaciones de un mismo sujeto estén en test o en training.
- Estratificación por la clase Disease o No Disease para mantener la proporción de casos tanto en el training y test con el desbalanceo de clases estable.
- División del conjunto total de datos disponible en subconjuntos (folds), de los cuales en diferentes iteraciones (Splits) utilizará uno de los fold como conjunto de test y la unión de los otros folds como conjunto de training. Entre las distintas iteraciones (splits) se garantiza que los conjuntos de tests no tienen sujetos repetidos.
- Promedio de los resultados de accuracy, AUC (y otras False positive, False negative) entre splits.

Se utiliza para tener estimaciones conservadoras («orientadas a tener idea del peor caso») de la capacidad de generalización de los modelos (media de los Accuracy, AUC etc) que complementen a las otras técnicas presentadas.

#### Consideraciones:

- Número de folds usado  $K=3$  (aprox. 70% training 30% test) para dar importancia al número de datos usados en el test y a la capacidad de generalización.
- Otros  $k$  mayores, por ej.  $K=10$  conducen a crear mayor número de folds con menor conjunto de datos en el test y dando más peso al training (pequeñas mejoras en accuracy y AUC).

### Comprobamos varios modelos con el dataset de voz (grande)

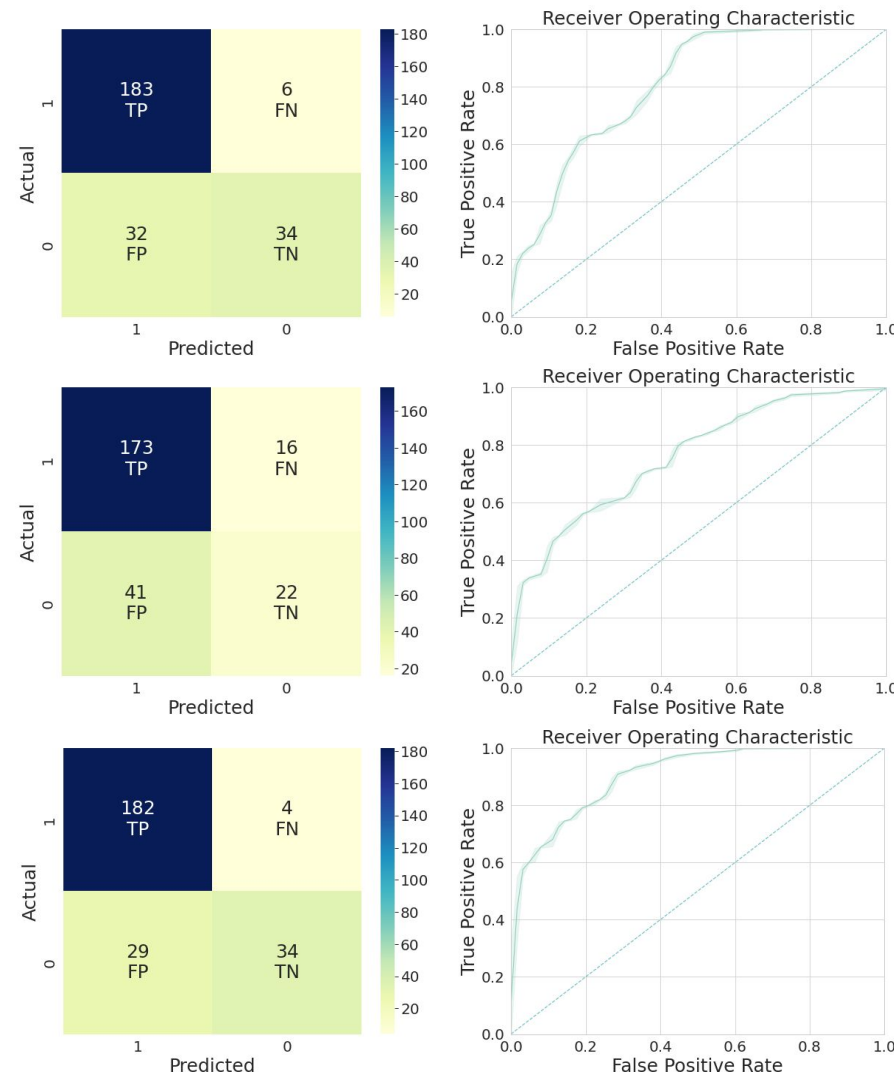
#### Accuracies (con StratifiedGroupKFold):

- KNN-3 = 0.7776470963217953
- SVM = 0.7594105647613798
- NB = 0.7007229665627964
- RF = 0.8109560557824201
- BC = 0.7884032863480064
- XGB = 0.8307532651614863
- CAT = 0.8227993205314253

#### Area under the curve (con StratifiedGroupKFold):

- KNN-3 = 0.6513393018769363
- SVM = 0.6670985326899306
- NB = 0.6635518006485749
- RF = 0.6706581948517433
- BC = 0.703414069005467
- XGB = 0.711022151882367
- CAT = 0.6956016902253461

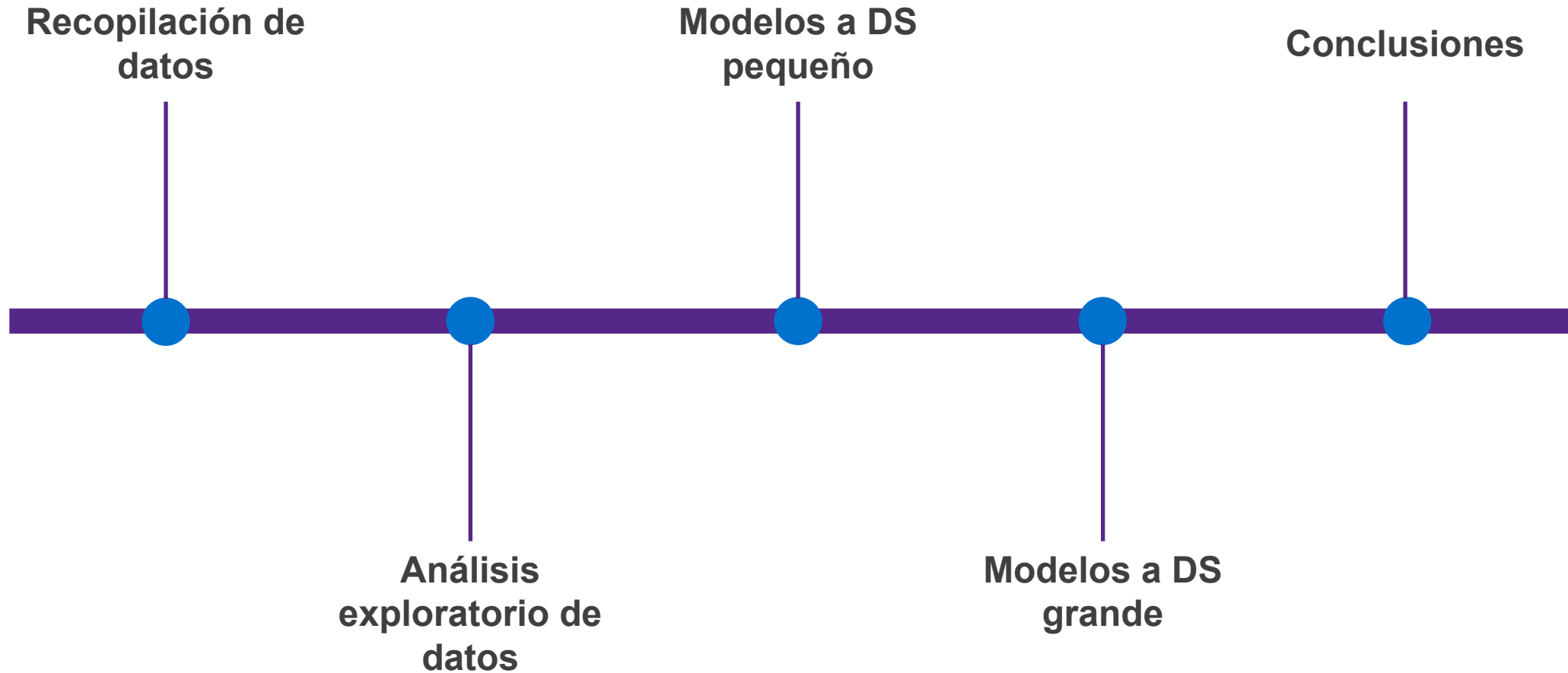
#### XGBoost



Split 1

Split 2

Split 3



# ¡Gracias por su atención!



UNIVERSITAT DE  
BARCELONA



DATA SCIENCE @ UNIVERSITAT DE BARCELONA