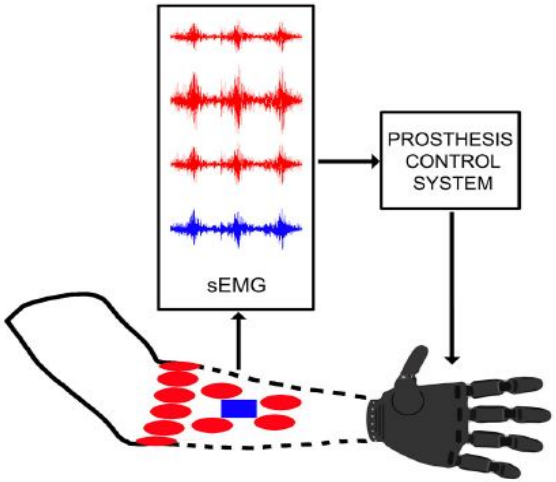
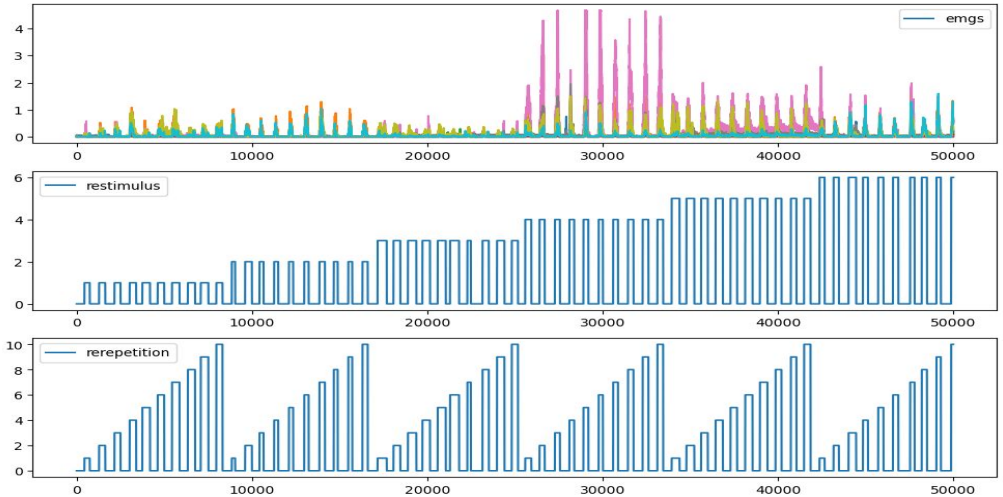
Aprendizaje de máquina 1

Jairo Agudelo M
Henry Arcila

Descripción del dataset

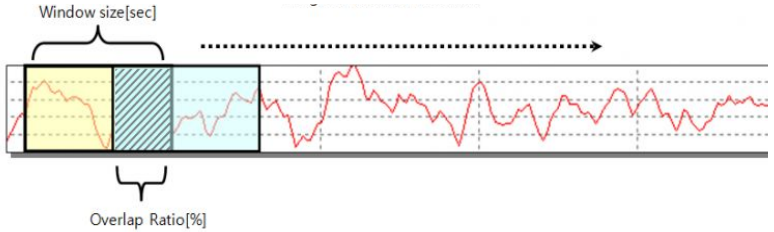
Ninapro DB1 (link)





Objetivo a desarrollar

Identificar, a partir de una muestra localizada de una señal electromiográfica superficial (ventana de tiempo), y usando modelos de clasificación, el tipo de postura de mano asociado a esta.

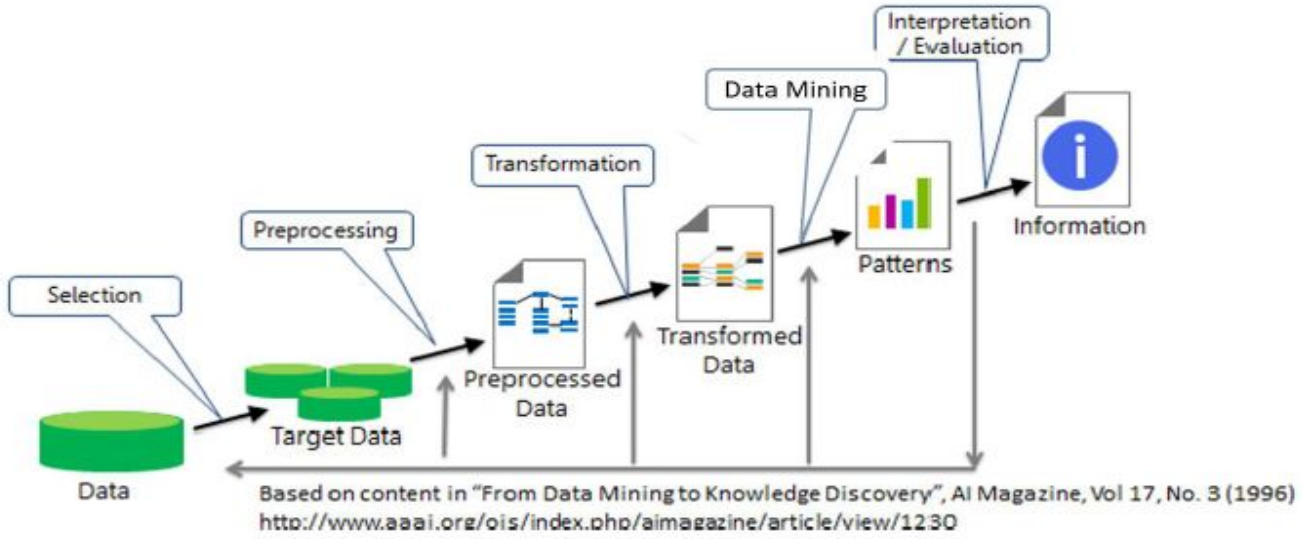


preprocesamiento de señales

Resumen del proceso realizado

La parte más difícil: el preprocesamiento de los datos

La parte entretenida: el ajuste de los parámetros de los métodos y la ejecución de los modelos

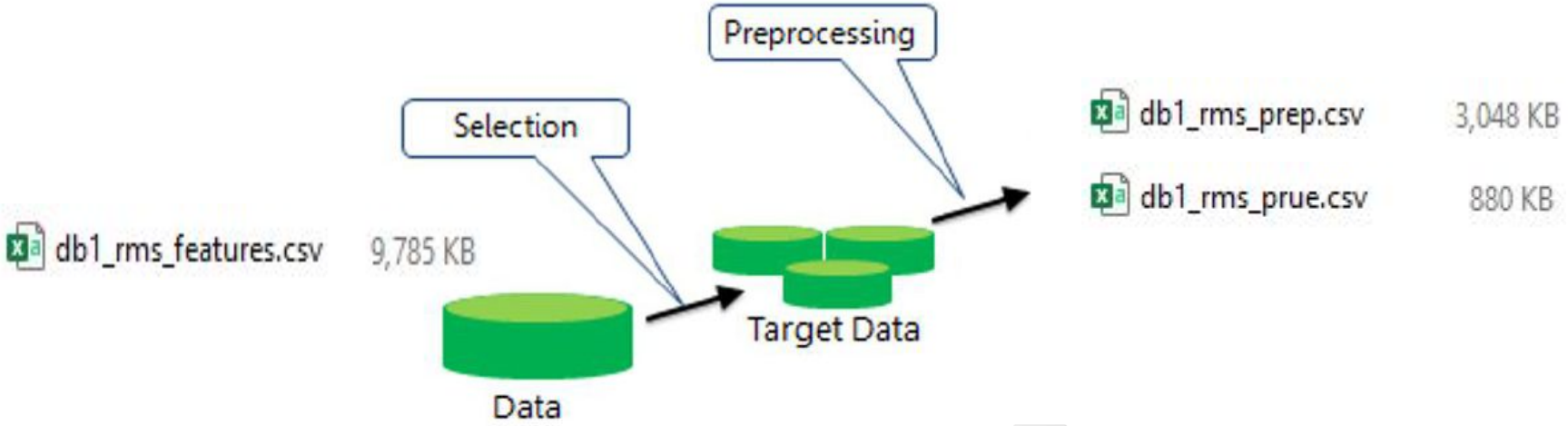


GitHub

link repo

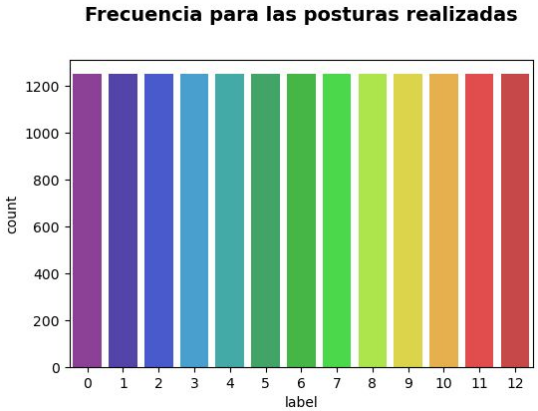
Desarrollo de experimentos

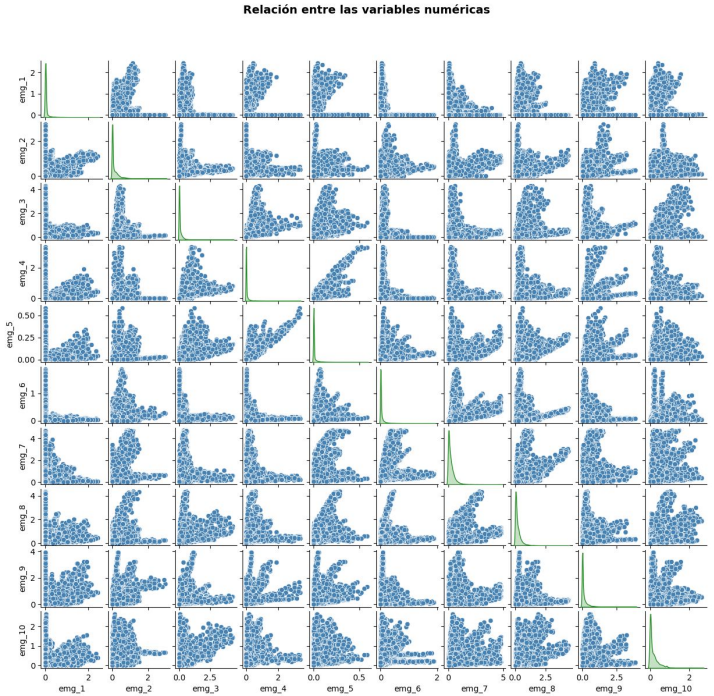
El inicio: selección de la base de datos y de los datos, el preprocesamiento de estos y la selección de los datos para el entrenamiento y para la validación de los modelos



preparacion de datos

Desarrollo de experimentos



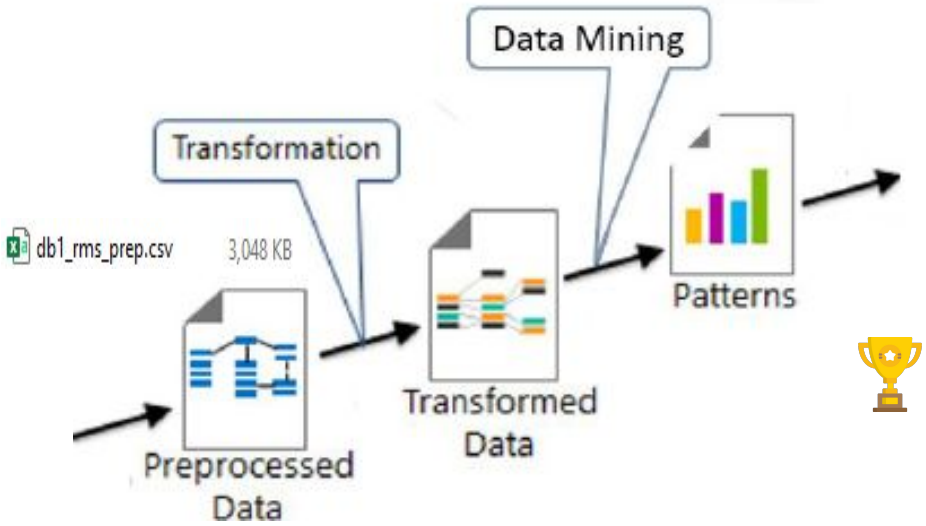


jupyter

[preparacion de datos](#)

Desarrollo de experimentos

El desarrollo: la transformación de los datos, la selección de los métodos y la creación de los modelos



Modelo 1: Regresión logistica ✓

Modelo 2: K-NN con Hiperparámetros ✓

Modelo 3: Árboles de Decisión ✓

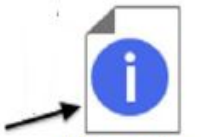
Modelo 4: Árboles de Decisión Random Forest ✓

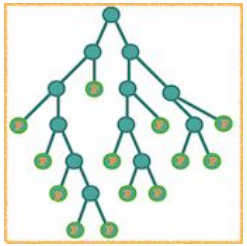
Modelo 5 - AdaBoost (Clasificación) ✓

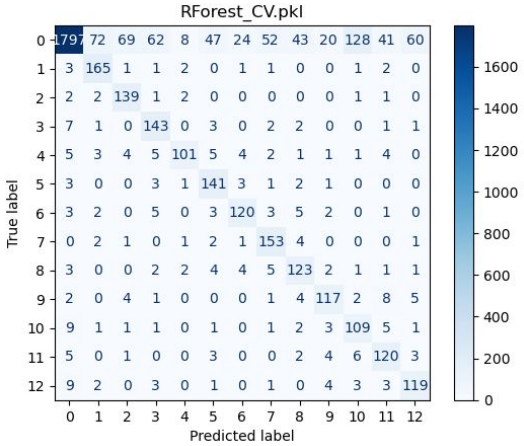
Desarrollo de experimentos

Evaluación de los modelos: el mejor al usar los datos de prueba fue Random Forest

AdaB_CV.pkl
DTreeC_CV.pkl
KNN_CV_manhattan.pkl
LR_Ret_ovrLineal.pkl
RForest_CV.pkl



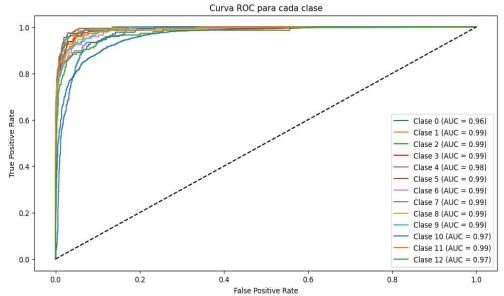




	precision	recall	f1-score	support
0	0.97	0.74	0.84	2423
1	0.66	0.93	0.77	177
2	0.63	0.94	0.76	148
3	0.63	0.89	0.74	160
4	0.86	0.74	0.80	136
5	0.67	0.91	0.77	155
6	0.76	0.83	0.80	144
7	0.69	0.93	0.79	165
8	0.65	0.83	0.73	148
9	0.76	0.81	0.79	144
10	0.43	0.81	0.56	134
11	0.64	0.83	0.73	144
12	0.62	0.82	0.71	145
accuracy			0.79	4223
macro avg	0.69	0.85	0.75	4223
weighted avg	0.84	0.79	0.80	4223

db1_rms_prue.csv 880 KB

Evaluación de modelos ✓



Conclusiones

- Como era de esperarse, la selección de los parámetros para los modelos es fundamental en los resultados de los mismos, para nuestro caso scoring f1 no arrojó resultados en 2 de ellos.
- La selección del número de pliegues(CV) también es fundamental ya que un número alto requiere mucho tiempo de computo y no necesariamente arroja mejores resultados en los modelos.
- Random Forest tardó cerca de 2 horas en correr los modelos con los parámetros que se seleccionaron.
- Los modelos basados en árboles son más susceptibles a los parámetros, valores bajos llevan a sub-entrenamiento "underfitting", con pocas opciones en la clasificación, y valores altos llevan a sobre-entrenamiento y "overfitting", aprendiendo las respuestas, perdiendo, en ambos casos, la capacidad de generalizar.
- A sabiendas de que la regresión logística es más adecuado para clasificaciones binarias quisimos probarlo en nuestro sistema multiclase arrojando como resultado una pobre clasificación.
- Ada-boost tampoco nos entregó buenos resultados ya que el clasificador simple en que se apoyó, árbol de decisión, mostró overfitting con los parámetro seleccionados. No insistimos en este modelo por los tiempos de ejecución altos.

Referencias

- Notas y notebooks del curso
- https://github.com/repos-especializacion-UdeA/trabajo-final_AA1