# Universidad de Buenos Aires
# Facultad de Ciencias Exactas y Naturales
# Facultad de Ingeniería
# Maestría en Explotación de Datos y Descubrimiento de Conocimiento
# Aprendizaje Automático

**1er cuatrimestre de 2020**

## Trabajo práctico Nro 2

El objetivo de este trabajo práctico es analizar las particularidades de la utilización de algoritmos de ensambles aplicados en casos casi reales. El mismo pretende fijar conceptos estudiados en la teoría: Naive Bayes; métodos de ensamble; random forests; boosting; sobreajuste; tolerancia al ruido. El material básico para la elaboración del presente trabajo se encuentra en las teóricas y prácticas presentadas hasta el momento y en las próximas clases y en las referencias bibliográficas indicadas [1,2,3]. Podrá utilizarse cualquier otra fuente siempre que esté correctamente referenciada.

El presente trabajo será grupal. El grupo deberá estar compuesto por exactamente cuatro integrantes asignados de forma aleatoria. Se evaluarán los contenidos del Trabajo Práctico durante el coloquio posterior a la entrega del TP junto al material teórico y práctico enseñado en las clases de toda la materia. Todos los integrantes deben tener conocimiento del desarrollo del TP.

La fecha límite de entrega es el 10 de julio a las 23:59.

Para el desarrollo del trabajo se utilizará el conjunto de datos Google Speech Commands Dataset (https://ai.googleblog.com/2017/08/launching-speech-commands-dataset.html) con el objetivo de predecir a partir del habla, qué digito se dijo.

Se deberá elaborar un informe preferentemente en LaTeX y entregarlo en formato .pdf. La entrega deberá estar acompañada de la Jupyter Notebook en Python utilizada para generar los resultados. El documento a entregar debe cumplir con los siguientes requisitos:

- debe tener no más que cuatro hojas, con fuente tamaño 10 e interlineado simple. La bibliografía no cuenta en la cantidad de hojas.
- una carátula en donde figuren universidad, nombre de maestría, materia, número de grupo, nombres de los integrantes del grupo, número de TP, año de cursada, etc. La carátula no cuenta en la cantidad de hojas.
- un resumen (del estilo de un artículo científico de no más de 200 palabras)
- una introducción en donde, entre otros, conste el objetivo del trabajo y una explicación de cómo está organizado el resto del documento.
- una sección de datos, en donde se describan los datos utilizados y sus particularidades

una sección de metodología, en donde se describan las metodologías utilizadas (sobre datos y sobre algoritmos)

una sección resultados, que incluya los resultados y su análisis

una sección de conclusiones. Por tratarse de un trabajo de investigación netamente práctico, las conclusiones deben ser la resultante de la elaboración de las pruebas realizadas. La información obtenida de referencias externas puede y debe ser tomada como insumo, pero no como conclusión.

referencias bibliográficas (referenciadas a lo largo del trabajo)

El informe se deberá publicar en el aula virtual de la materia por uno sólo de los integrantes del grupo.

Para realizar el informe se deberán considerar y documentar los siguientes puntos:

a) Extraer atributos a partir de los audios correspondientes a los digitos del 0 al 9. Utilizar el material suplementario dado durante la clase de presentación del TP.
b) Utilizar la división de datos provista en el dataset para entrenamiento, validación y evaluación de los modelos a desarrollar.
c) Entrenar modelos de Naive Bayes, Random Forest y Gradient Boosting Machines para predecir a partir de los atributos acústicos el dígito pronunciado.
d) Evaluar y comparar el rendimiento (performance) de los modelos. Se podrá elegir la medida de rendimiento que se estime más adecuada. Deberá justificarse la elección. Reportar la matriz de confusión de alguno de los modelos.
e) Analizar los errores cometidos por los modelos.
f) Evaluar el impacto en el rendimiento de agregar distintos niveles de ruido en los audios de evaluación. Utilizar como señales de ruido:
   i) Ruido gaussiano.
   ii) Audios de ruido ambiental.
   Algunos datasets con muchos audios de ruido ambiente son:
   https://github.com/qutsaivt/QUT-NOISE
   https://www.kaggle.com/aanhari/demand-dataset
   https://urbansounddataset.weebly.com/urbansound8k.html

g) Grabar audios pronunciando los dígitos en distintas condiciones ambientales. Por ejemplo, se puede probar grabando en distintos ambientes (baño vs. cocina), distintas distancias al micrófono, pronunciando de distintas maneras y utilizando distintos dispositivos (celular vs. laptop). Analizar los resultados y el efecto de las distintas condiciones sobre la performance de los modelos.

Opcional (da puntos extra): implementar un modelo de perceptrón multicapa. Elegir el número de neuronas de la capa oculta utilizando random search. Comparar su rendimiento con los demás modelos desarrollados en c).

**Referencias**

[1] An Introduction to Statistical Learning. Capítulos 2 (2.2.2), 5 (5.2) y 8 (8.2, 8.3.3, 8.3.4). http://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf

[2] Seni, Elder, “Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions”, Morgan & Claypool, 2010. https://doc.lagout.org/Others/Data Mining/Ensemble Methods in Data Mining_ Improving Accuracy through Combining Predictions %5BSeni %26 Elder 2010-02-24%5D.pdf

[3] http://scott.fortmann-roe.com/docs/BiasVariance.html