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 2019

Trabajo práctico Nro 2

Fecha de entrega digital: 21 julio 23:59 hs.

Fecha de entrega impreso: lunes 22 y jueves 25 (dependiendo de la comisión)

El objetivo de este trabajo práctico es construir modelos clasificadores basados en ensambles.

Para el trabajo cada grupo podrá utilizar los conjuntos de datos que utilizó en el trabajo práctico anterior o bien el siguiente dataset de Kaggle, que se podrá obtener del siguiente link: https://www.kaggle.com/danimal/heartdiseaseensembleclassifier. El dataset contiene datos de pacientes respecto a la aparición de enfermedades cardíacas. El dataset tiene 14 atributos y 303 observaciones. Entre los atributos se incluyen factores de riesgo cardíaco, como ser sexo, edad, nivel de colesterol entre otros.

Se pide

1. A partir del dataset se deben elaborar tres modelos de predicción con distintos métodos de ensamble.

2. Se debe informar para cada modelo, TP, FP, TN y FN, Accuracy, Precision, Recall y F1.

3. Utilizar 5-fold cross validation habiendo separado anteriormente 20% del dataset para test (de manera aleatoria). Informar cuál es el modelo que mejor performance, demuestre en el conjunto de test.

Entregar (digitalmente subiéndolo al aula virtual):

• Un informe (cuyo archivo se deberá denominar “tp2-Grupoxx.pdf”, siendo xx el número de grupo) en los términos que se especifican más abajo. El informe debe entregarse impreso también.

• Un archivo con el mismo nombre pero extensión “.txt” con los datos del dataset.

Otros detalles

El grupo deberá estar compuesto por los mismos tres integrantes que en el primer TP. Si algún grupo quedó con menos de tres integrantes, debe rearmarse.

Todos los integrantes deben tener conocimiento del desarrollo del TP.

El trabajo deberá implementarse en Python.

Informe

El documento a entregar debe cumplir con los siguientes requisitos:

- una carátula en donde esté el nro. del grupo, sus integrantes, nombre de maestría, materia, etc.

- un resumen (del estilo de un artículo científico)

-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 metodologías, 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 y trabajos futuros. 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)

Se deberá entregar digitalmente el informe, el dataset, el código ejecutable y se deberá entregar el informe impreso.

Sin contar el código el informe no deberá tener más de 6 páginas a espacio simple en Arial 11 y se deberá publicar en el aula virtual de la materia por uno sólo de los integrantes del grupo.