**Actividad Big Data**

**Fecha de entrega: mayo 9 de 2020**

**José Fernando Zea**

La siguiente actividad pueden trabajarla en grupos de a tres (con excepción del punto 1).

1. Individualmente Realice la ejecución de los scipts <https://spark.rstudio.com/mlib/>

No ponga enfasis en los detalles técnicos de la técnica de kmeans, enfatice la regresión, regresión logística, random forest y los procesos para seleccionar una muestra de entrenamiento y de prueba. Este punto no se entrega.

1. Utilice *databrics* para usar la base unificada de las pruebas saber 2019 (2019I y 2019II integradas por debajo), sin seleccionar datos de entrenamiento y de test, lleve acabo los siguientes modelos:
2. Una regresión lineal múltiple con variable dependiente el puntaje en matemáticas. Use como variable predictiva solo dos asignaturas (las dos que vea más correlacionada), puede hacer uso de las demás variables presentes en la base de datos (en su gran mayoría variables categóricas). Seleccione previamente las variables a utilizar y asegurese que no tenga registros faltantes en las variables que va a utilizar para su modelo. Esocoja su modelo más apropiada y calcule el mse.
3. Una regresión random forest con variable dependiente el puntaje en matemáticas. Use como variable predictiva solo dos asignaturas (las dos que vea más correlacionada), puede hacer uso de las demás variables presentes en la base de datos (en su gran mayoría variables categóricas). Seleccione previamente las variables a utilizar y asegurese que no tenga registros faltantes en las variables que va a utilizar para su modelo. Esocoja su modelo más apropiada y calcule el mse. Compare con el modelo del punto a.
4. Una regresión logística donde la variable a predecir es el desempeño en inglés (considere dos clases, los estudiantes que obtuvieron A y los estudiantes que obtuvieron B). Puede escoger al menos tres asignaturas y las demás variables a nivel de estudiante o de institución que considere. Calcule la matriz de confusión, la precisión, sensibilidad y especificidad e investigue como obtener una métrica adicional
5. Repite el punto 2 pero está bien haciendo la selección de una muestra de entrenamiento y de test.

Nota: Crossvalidation (tema de la siguiente clase). No realizar por ahora.