# Plan de ejecución para la segunda iteración de las etapas

Juan Carlos López López  Adolfo Marín Arriaga  Luis Rodrigo Rojo Morales

3 de diciembre de 2016

En la primera iteración lo que nos ayudo bastante para hacer preprocesamiento de los datos fue graficarlos, estas gráficas la podemos ver en el archivo /Entregable2/Preprocesamiento/Preprocesamiento.pdf, después de ver el comportamiento de los datos, vimos que algunos datos numéricos estaban muy variados, por ejemplo en *hours_per_week* los datos van desde 0 horas hasta 100 horas, pero mas de la mitad estaba en 40 horas, por lo que se nos hizo buena idea aplicar *Binning*, con esta técnica nos quedaron rangos de 0 horas a 20 horas, 21 horas 40 horas,..., 81 horas a 100 horas. Y análogamente para el atributo *age*.

También los atributos *capital_gain, capital_loss y native_country* cada uno tiene un valor el cual es el más frecuente, por lo que decidimos solo enfocarnos en ese valor al momento de hacer las tablas de frecuencias, las cuales son necesarias al usar el algoritmo Naive Bayes, por ejemplo, en *native_country* teníamos que el atributo más frecuente era *United-States*, entonces todos los demas valores distintos a este le pusimos *Foreigner*, similarmente con *capital_gain, capital_loss* pero con el valor 0, esto nos ayudo bastante a bajar la carnalidad de la tabla de frecuencias.

Al momento de hacer la mineria de datos en particular Naive Bayes la implementacion resulto fácil, puesto que solo habia que tener la tabla de frecuencias y aplicar las formulas correspondientes, mientras que para implementar C4.5 ya no resultó tan facil porque tuvimos problemas al general el árbol

Hicimos cambios en el preprocesamiento los cuales mejoraron 1 % al resultado de >50k de Naive Bayes, pero empeoraron la prediccion de $\leq$ 50k