Informe de la cita, día 2017.10.26:

En la cita, presenté a la tutora un informe del trabajo que había hecho en las tres semanas transcurridas, que se dividía básicamente en dos partes: 1. el trabajo preparatorio, que consistía en preparar el entorno para estudiar el problema y para desarrollar el proyecto, también se incluían la formación previa y la preparación del código; 2. luego es la implementación de funcionalidades, aquí le expliqué el resumen del desarrollo (el código implementado hasta ese momento), así como las ideas de atrás pensadas, también saqué algunos resultados: dimensión de los datos, tipos de predictores empleados, presentación de los outputs de predicciones, comparación de performances de cada tipo de predicción. Cabían destacar las dos aproximaciones que se habían tenido en cuenta a la hora de predecir las notas: si primero se predice con las notas o directamente se clasifica con el ranking (o estado binario, para la clasificación de suspensos).

No obstante, los resultados obtenidos no eran del todo correcto por los siguientes problemas detectados:

1. Remplazo de los valores “NaN” por 5.0, no es razonable.
2. La función usada para calcular la puntuación del “10-fold cross validation” podría no ser correcta y hay que revisarla.
3. La medida de validación usada para cada predicción podría no ser adecuada, hay que buscar una adecuada de forma que los resultados sean comparables.
4. Hay que averiguar si el RF testea de forma independiente o conjunta las clases existentes. Si es el primer caso, diríamos que no se está clasificando bien, ya que se espera que entre las clases predichas haya una relación.
5. Se observa que cuando se clasifica por el ranking, se da un resultado en que aparecen clases repetidas, por eso se puede decir que no es una clasificación estricta, la que necesitamos. Para ello la tutora propuso hacer un recomendador, el cual puede realizar la clasificación estricta, también solucionaría el problema 4.

Por todo lo dicho anteriormente, no podemos confiar en los resultados mencionados (outputs y performances), al menos la mayor parte. Tampoco podemos concluir ahora cuál de las dos aproximaciones es mejor (predice con mejor precisión), antes de que todo sea revisado.

Para la siguiente cita, las tareas son:

1. Solucionar los problemas detectados, los 5 puntos anteriores. Para el punto 5, hay que mirar cómo están implementados los recomendadores en el código de Sergi.
2. Si diera tiempo, procesar datos de “Informática”.