Informe de la cita, día 2017.10.04:

Al inicio, expliqué a la tutora lo que había hecho en las tres semanas anteriores, básicamente era mirar el informe del Sergi y el código implementado por él, también aprendía cosas relacionadas con el aprendizaje automático, mediante vídeos en Youtube, ya que creo que me hace mucha falta.

Antes creí que mi TFG se consiste en hacer un proyecto paralelo a los de los previos alumnos, o sea, implementar las mismas funcionalidades, ya sea predicción de la tasa de abandono, o de las calificaciones con diferentes algoritmos de clasificación y diferentes datos (notas curriculares de alumnos de otras facultades). Me equivoqué, pues la tutora dijo que yo tendría que hacer cosas nuevas, la idea que propuso la tutora era una variación del "ranking de asignaturas del segundo curso" para los alumnos del primer curso, esta funcionalidad la había implementado el "Sergi Rovira" en su TFG 2016-17. A la parte conclusiva del mismo TFG, aconsejó mejorar "predicción de notas", y de aquí surgieron dos consejos: primero, en vez de hacer un ranking de asignaturas por calificaciones que obtendría un alumno, se haría un ranking por número ordinal; segundo, hacer un clasificador de suspensos.

Descripción de las funcionalidades nuevas:

1. Ranking de asignaturas por número ordinal: mediante la información de las calificaciones de las 10 asignaturas de los alumnos del primer año, se predirán las 10 que un alumno podría sacar en el segundo año, y se hará un ranking ordinal de éstas, del cuál el número 1 es la asignatura con calificación más alta, el número 2 es la segunda más alta, y así sucesivamente.
2. Clasificador de suspensos: mediante la información de las calificaciones de las 10 asignaturas de los alumnos del primer año, se predirán las asignaturas que un alumno podría suspender, es decir, con una calificación menor que 5.

Descripción de la implementación:

1. Selección de algoritmo: la tutora propuso usar el "Random Forest", un clasificador múltiple, ya que en el primer problema se clasificarán los resultados en 10 clases (del número 1 a 10). También se utiliza para el segundo problema.
2. Obtención, procesamiento de datos y limpieza de datos: en este TFG no hace falta realizar estos procesos relacionados con datos, como se ha mencionado anteriormente, en este TFG se realizan posibles ampliaciones y mejoras basadas en el proyecto del dicho alumno (Sergi Rovira), por lo cuál la implementación de datos será reutilizada.

3. Validación del algoritmo: la tutora propuso usar el “10-Fold Cross Validation”. Es una validación en la que se divide el dataset en “training data” y “testing data”, la proporción es 90% y 10%, respectivamente. Después de la división, se hace una predicción y se calcula el “accuracy” para cada uno de los 10 turnos, al final se calcula el promedio. Véase la figura 1

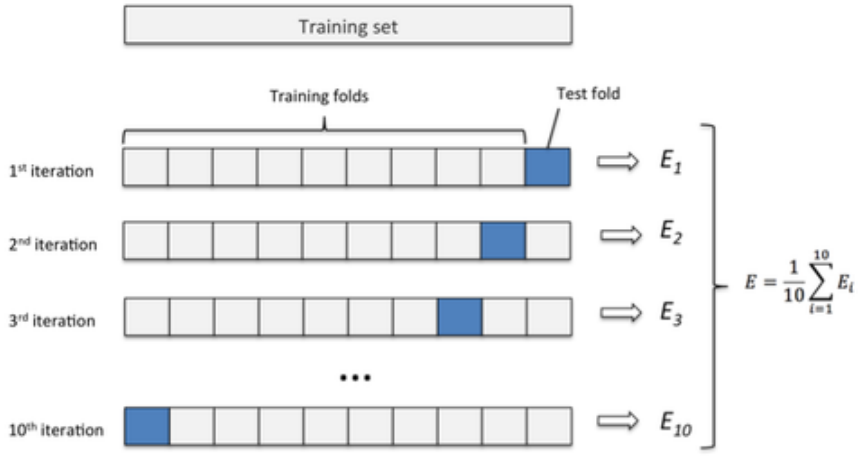

Fig 1: validación cruzada de 10 iteraciones

La tutora advirtió traer el resultado de las pruebas realizadas y las dudas para cada cita, así como el powerpoint de la explicación del avance, también hay que escribir un informe después de cada cita.

Antes de la próxima cita día 2017.10.18 a las 16.00h, seguiré mirando tutoriales, repasaré el código del Sergi e implementar el Random Forest para los dos problemas y si es posible, sacar el resultado del mismo.