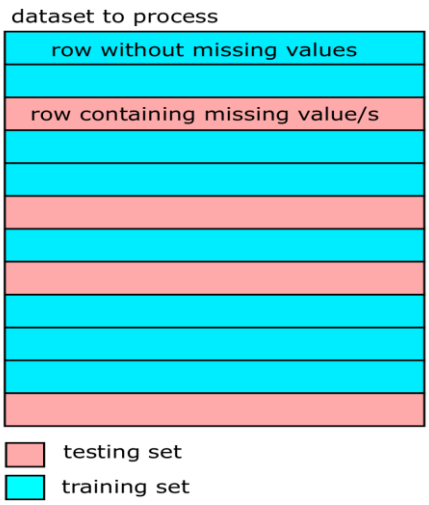
Informe del progreso de TFG, fase 3 (2017.11.16-2017.11.30)

1. *Implementación de **ponderación de puntuaciones de similitud de Pearson (PPP),** los métodos implicados son:*
    a. KNN: se calculan las similitudes entre el alumno en testeo y los alumnos del training set, y los K alumnos con mayor puntuación serán escogidos.
    b. recommend: se ponderan las notas de los K alumnos por sus similitudes correspondientes al alumno en testeo, la puntuación ponderada será el resultado de predicción.

2. *Casos de **MV** en el dataset inicial:*
    a. Mantener los MV: este caso no servirá para la aproximación de "predicción mediante ranking", ya que no se pueden rankear las notas cuando hay MV. Para la "predicción mediante calificaciones", se han tenido en cuenta los posibles problemas de MV, por tanto se han puesto diferentes controles (filtros, condiciones).
    b. Eliminación de los MV: antes de la predicción, se eliminan las filas en que hay MV.
    c. Reemplazo de los MV: antes de la predicción, se reemplazan los MV mediante "RecMvRepl". Este caso presenta un contraste al caso anterior, porque pretende mantener la dimensión del dataset inicial.

3. *El problema de la asunción de MV:*
   La asunción de MV como pésimos rankings no está funcionando bien, cuando hay MV en los datos, los datos no se ordenan bien (de forma esperada, se pensaba al hacer la ordenación, los MV son menores que los demás). Se decide **abandonar** esta opción.

4. *La clase "RecMvRepl":*
   Recomendador que reemplaza los MV, explicada en palabras. Es una clase que se encarga de recomendar valores de predicción en posiciones de MV. Tiene la misma estructura de un recomendador normal, pero el mismo dataset es utilizado como X (vector de características) e y (etiquetas).



Es decir, el dataset se divide en training set y testing set:
el **training set** está formado por las filas que no contienen ningún NaN, y es usado para el entrenamiento, y predecir;

el **testing set** está formado por las filas que contienen al menos un NaN, y es usado para el entrenamiento, el resultado recae en el mismo (reemplazo de NaNs)

5. Limpieza de datos: antes se eliminan las filas del dataset concatenado (X e y) en que el número de valores non-NaN no llega a 11 (umbral=11). Pero una vez aplicada este filtro, se observan filas con meramente uno o dos valores non-NaN en el dataset de **y**, estas filas son poco significativas para el proceso de predicción. Por tanto, se ha modificado la manera de filtración: se observan **X** e **y** separadamente, si alguna fila contiene **más de 5 NaNs**, será eliminada de **X** y a la vez, la misma fila (según índice) será eliminada de **y**.

6. Tareas para la siguiente cita:
    a. Rellenar las dos tablas de evaluación
    b. Testear con el dataset de “Informática”
    c. Estudiar el RandomForestRegressor y RandomForestClassifier (Si da tiempo)