**INFORME PARA LA SIGUIENTE REUNIÓN**

1. Se ha probado a transformar la variable CONSTRUCTIONYEAR a una ANTIQUITY, de año de construcción a antigüedad “ANTIQUITY = 2020 – CONSTRUCTIONYEAR”. Las métricas de *accuracy* y *recall* son las mismas para ambos modelos (XGBoost y Random Forest) que las de la entrega de la Fase I.
2. En la variable MAXBUILDINGFLOOR, cuando era un valor nulo, se transformaba a -1 y ahora a 0. Las métricas de *accuracy* y *recall* son las mismas para ambos modelos (XGBoost y Random Forest) que las de la entrega de la Fase I.

CONCLUSIÓN: si estos cambios mejoran o empeoran la precisión muy ligeramente, es prácticamente imposible detectarlo.

1. Se ha probado a transformar la variable CADASTRALQUALITY de tipo categórica a one-hot. Para ello se ha eliminado esta y añadido las 13 one-hot correspondientes. Se ha generado un fichero respuesta con 100 iteraciones con este cambio y los dos anteriores. Dicho fichero es muy similar al de la entrega de la fase I, aunque contiene menos muestras predichas a la clase RESIDENTIAL.

CONCLUSIÓN: Posiblemente estos cambios mejores muy ligeramente la precisión.

1. Se ha entrenado un modelo Support Vector Machine sin realizar PCA a los datos, pero con los cambios anteriores realizados. La precisión se reduce al 23% y el *recall* al 14%.
2. Se están haciendo pruebas con MLBox (técnica de auto machine learning) tanto en un ordenador personal como en Google Collab pero de momento no muestra nada (parece estar bloqueado)
3. Se ha probado el modelo Extra Trees (con entre 200 – 400 estimadores), da resultados ligeramente peores que el Random Forest (tanto en *accuracy* como en *recall*)
4. Se ha realizado Grid Search para Random Forest y XGBoost pero nos encontramos con un problema. Estos algoritmos esperan un conjunto de datos equilibrados (cosa que realizamos mediante *under sampling* de RESIDENTIAL) y tras lanzar varias ejecuciones, los hiperparametros que nos devuelven son diferentes y distan entre sí. Esto es debido a que como las muestras añadidas de RESIDENTIAL son diferentes cada vez la búsqueda que hace cambia. Nos hemos quedado con una estimación de los hiperparametros que más salían ganadores.

CONCLUSIÓN: n\_estimators = 750, criterion = ‘gini’,

1. Se ha entrenado a un modelo Random Forest con los datos tras aplicarles SMOTE y otras técnicas de over i under sampling. Se ha conseguido una precisión de 96% pero esta valor es de suponer que ha sufrido overfitting.