Trabajo Práctico Final: Problema real de aprendizaje automático y ciencia de datos

Objetivo:

El objetivo de este trabajo práctico es enfrentar al alumno a un problema real de ciencia de datos utilizando los métodos de aprendizaje automático vistos en la materia. Para ello utilizaremos un conjunto de datos de prueba, obtenido de Kaggle y el cual es parte de un problema real propuesto por el banco Santander. Dicho banco se pregunta: “¿Qué clientes harán una transacción especifica en el futuro sin importar la cantidad de dinero de la transacción?”.

Datos:

Por ello provee una serie de datos en formato CSV correctamente ofuscados, lo cual convierte al problema en un simple problema de clasificación. En los archivos de entrenamiento encontraremos 1 variable dependiente con 2 valores posibles 0 y 1 (si el cliente hará o no hará dicha transacción) y 200 variables numéricas independientes.

> **Descarga**: https://drive.google.com/file/d/1xHXxmHV8eJmSzExkB-RMZjl1fX6tTa90/view?usp=sharing

Desarrollo del trabajo:

Dividir el conjunto de datos en entrenamiento y prueba, usar el criterio de ⅔ para entrenar y ⅓ para probar.

El trabajo consiste en realizar los siguientes pasos.

1. **Exploración de datos**. No hay demasiado que hacer que los datos están ofuscados, aunque quizá algún que otro dato se podrá deducir y considerarlo.

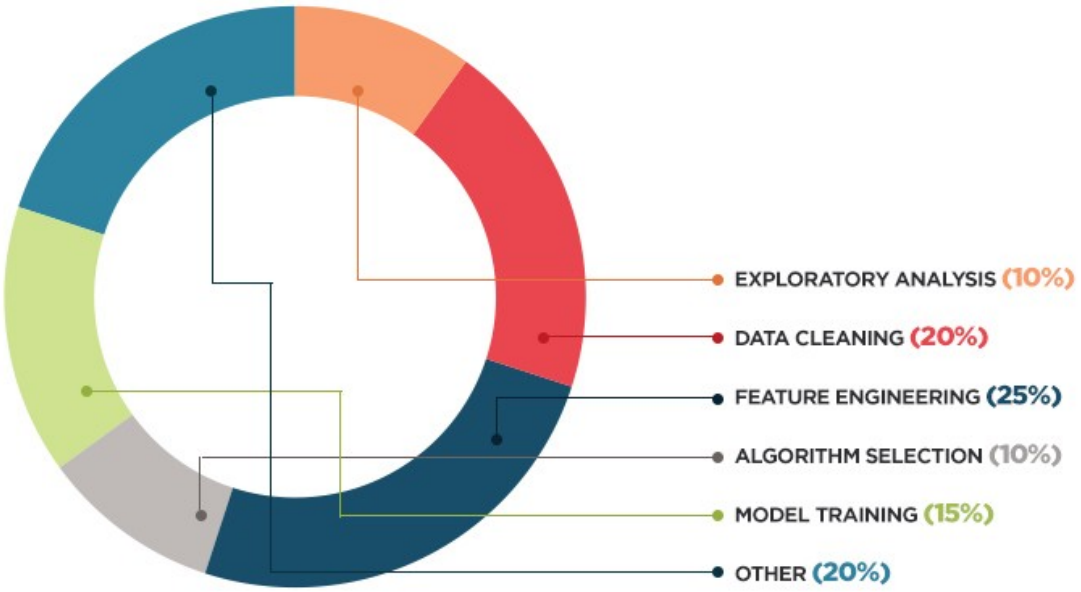
2. **Limpieza de datos**: quitar valores atípicos (*outliers*) y eliminar casos con datos faltantes o bien completarlos. Tampoco hay demasiado que hacer ya que esta parte la realizó el banco Santander en principio.

3. ***Feature engineering****:* probar los valores con el modelo elegido, ver si es posible crear *features* nuevos de existentes, si de un valor se extraen más de uno, mejorarlos, etc. Usar ademas alguna herramienta de *feature engineering* si es posible.

4. **Elegir algoritmo**: de los vistos en clase: Arboles: ID3, C4.5, *Random Forest;* Redes neuronales, regresión logística, *Bayes Naive*. U otros que el alumno considere. Tener en cuenta la posibilidad de usar más de un método combinado. También considerar la posibilidad de usar un método de

agrupamiento (*clustering*) como SOM (Kohonen), K-Means, etc. para subdividir el conjunto en subconjuntos y entrenar cada subconjunto por separado.

5. **Probar el modelo**: Usar el modelo de prueba, para probar que tan bien funcionó la predicción. Para ello calcular la precisión, el *recall* y la medida F1.



Este ciclo se puede ejecutar a mano o utilizando meta-algoritmos como los vistos en clase que automáticamente prueban varios modelos. Es la recomendación de la cátedra.

Probar con al menos 3 métodos.

Forma de entrega: El trabajo se deberá entregar por e-mail y deberá incluir:

- El archivo con el código fuente o *link* al repositorio o acceso a la maquina colab. Lo que el alumno prefiera.
- Un informe en PDF en formato LNCS. LNCS es un formato para la publicación de artículos científicos. Y es el que utiliza CACIC que es el congreso de ciencias de la computación más importante de la Argentina. Les pido este formato para que se vayan familiarizando con este tipo de documentos ya que los artículos científicos son los documentos a los cuales tendrán que acceder cuando busquen soluciones a problemas originales, aún no resueltos como el que se propone en este trabajo.

Les paso a continuación los *links* de descarga a los *templates* para Word y *Latex* de dicho formato. Daremos más detalle sobre los artículos científicos y su formato en clase.

- https://drive.google.com/file/d/0B3T3L0K6Tm8RanFUeEVERDM2VXM/view?usp=sharing
- https://drive.google.com/file/d/0B3T3L0K6Tm8RUXlDVGVuZGYyLTg/view?usp=sharing
- https://drive.google.com/file/d/0B3T3L0K6Tm8RUXlDVGVuZGYyLTg/view?usp=sharing

Contenido del Informe:

- Qué tareas de limpieza de datos se realizó.
- Qué tareas de *future engineering*
- Qué algoritmos se seleccionaron
- Qué resultados arrojó cada algoritmo en el conjunto de pruebas
- Qué método finalmente se seleccionó (o combinación de métodos) y cual fue el resultado en el conjunto de pruebas
- **Conclusiones**: ¿Se puede o no se puede predecir? Si no, ¿Qué falta? , etc.

Fecha de entrega: El trabajo practico final puede ser enviado en cualquier momento hasta dos semestres después de terminada la cursada. Al igual que el tiempo que se cuenta para dar exámenes finales, mientras no se venza la materia.