**Proyecto de Machine Learning**

**Problema:**

Dado un chat saber si en algún momento uno de los usuarios comienza a escribir diferente ya sea por una razón u otra.

Como sub-problema se tiene que dado un texto saber cuál de los integrantes del chat escribió dicho texto.

**Modelación del problema:**

Dado un chat se seleccionaron una cantidad de mensajes que comienzan en algún momento. En un diccionario se guarda por cada usuario todos sus mensajes. Haciendo uso de la biblioteca nltk de Python se realiza un procesamiento de texto donde por cada mensaje se va a eliminar las stopwords, o sea, las palabras que no aportan ningún significado, los signos de puntuación y se lematiza el texto, se llevan las palabras a su raíz gramatical. Luego de este procesamiento se vectorizará cada mensaje usando el TfidfVectorizer de la biblioteca sklearn y así se obtendrán los vectores de cada usuario por cada mensaje.

**Algoritmos de aprendizaje no supervisado usados en la resolución del problema:**

**- One Class SVM:**

Aprende una función de decisión para la detección de novedades: clasificar nuevos datos como similares o diferentes al conjunto de entrenamiento.

**- Robust Covariance:**

Detecta valores atípicos en un conjunto de datos distribuidos de Gauss.

**- Isolation Forest:**

Detecta anomalías mediante el aislamiento, en lugar de modelar los puntos normales.

**Algoritmos de aprendizaje supervisado usados en la resolución del problema:**

**-KNN**

**-Naive-Bayes**

**-Decision Tree**

**-Random Forest**

Estos algoritmos fueron los que estudiamos en clase. Escogimos KNN y Naive-Bayes porque trabajamos un problema similar en la CP2, aparte de que la idea de cercanía entre los conjuntos que tienen ambos algoritmos se asemejan a la idea que utiliza en un principio Novelty Detection, aunque enfocado en el aprendizaje supervisado por supuesto.

Por otro lado, utilizamos los modelos de Decision Tree y Random Forest porque demostraron ser de los modelos que mejores resultados obtenían en jeugos de datos de este estilo, siempre que se cuidara bastante el no realizar overfitting.

**Análisis de los resultados:**