**Para el desarrollo de este proyecto nos apoyaremos del proceso KDD Knowledge Discovery in Databases que básicamente contiene fases ya definidas que nos van a permitir descubrir conocimiento a partir de una fuente de datos**

**1 Fase selección de fuente de datos**

Para el desarrollo del proyecto se uso como fuente de datos el data set uci-har que se encuentra en el repositorio <https://archive.ics.uci.edu/dataset/240/human+activity+recognition+using+smartphones> , el cual contiene informacion de registros de 30 sujetos de prueba que realizaban actividades diarias mientras llevaban un teléfono inteligente montado en la cintura con sensores inerciales incorporados a saber : Acelerómetro y Giroscopio , el registro de las señales emitidas por estos dispositivos se dio en las dimensiones o ejes X;Y;Z.

Se enfoco básicamente en 6 actividades que debian realizar las personas : CAMINAR, SUBIR, BAJAR, SENTAR, DE PIE, TUMBADO , acciones que eran medidas mediante el acelerómetro y giroscopio incorporado en el dispositivo móvil , se registraron aceleración lineal y velocidad angular triaxiales a una frecuencia constante de 50HZ , los datos obtenidos se dividieron en dos conjuntos de datos de forma aleatoria ,70% para datos de entrenamiento y 30% para datos de prueba.

Las señales de los sensores se preprocesaron mediante la aplicación de filtros de ruido y se muestrearon en ventanas deslizantes de ancho fijo de 2.56 segundos y un solapamiento de 128 lecturas por ventana .La señal de aceleración del sensor que incluye componentes gravitacionales y de movimiento corporal se separo mediante un filtro de paso bajo Butterworth

De lo anteriormente expuesto se desprende que el dataset utilizado para este proyecto ya se encuentra dividido en conjuntos de entrenamiento y test ; también podemos constatar que dicho data set contiene dos versiones de dataset , el primero corresponde a los registros preprocesados , donde cada columna es una feauture procesada y contiene 561 columnas por fila , normalizadas y listas la aplicación de modelos como RandomForest , SVM, KNN y el segundo contiene las señales crudas de los sensores antes de ser transformadas en características , que contiene series temporales de Aceleración del cuerpo, Aceleración por gravedad y Velocidad angular .

Dado que la primera versión contiene los datos ya procesados y separados para entrenamiento y test , utilizaremos esta primera versión para entrenar un modelo de clasificación Random Forest Classifier , para el resto de modelos a ser abordados en este proyecto se hará uso de la segunda versión del dataset , la cual contiene señales crudas de los dispositivos.

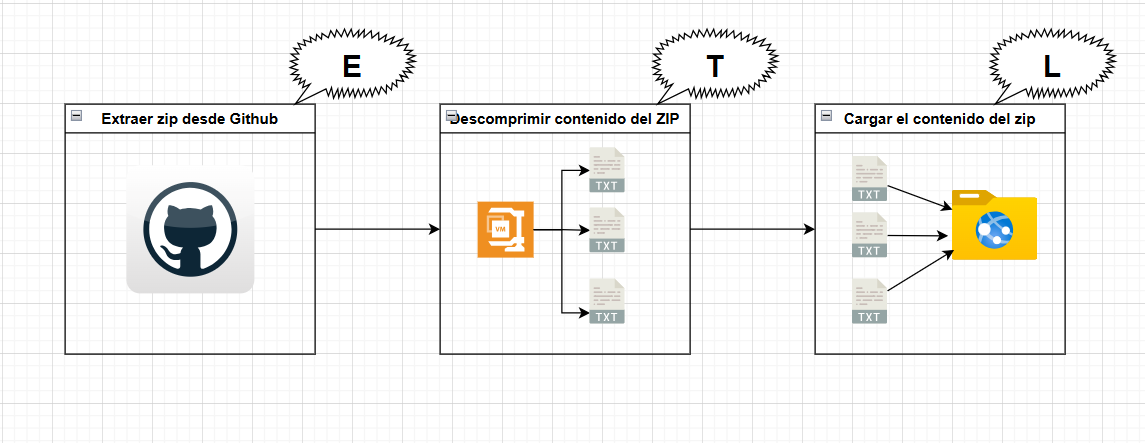
* 1. Proceso ETL para carga del dataset desde GITHUB

Para este proyecto el dataset de uci-har se encuentra comprimido en formato zip , dentro del cual se encuentran las dos versiones de los datase a ser utilizadas(procesados, datos crudos) , en virtud de esto se procede a generar un proceso ETL básico que :

Extraiga el comprimido uci-har.zip que se encuentra en un repositorio github

Descomprima el archivo zip para poder utilizarlo

Cargar el contenido del archivo zip en un repositorio local con el propósito de que sirva como insumo para el análisis de datos.



Para la construcción de este ETL se utilizo como lenguaje de programación Python y las siguientes librerías :

|  |  |
| --- | --- |
| zipfile | Se utiliza para leer, comprimir y descomprimir archivos zip |
| io | Contiene métodos que nos permiten crear objetos tipo archivo en memoria |

,

**Fase de preprocesamiento**

Una vez identificado el insumo de datos con el cual se va a laborar procedemos a ejecutar el análisis exploratorio de datos EDA , como punto de partida para el análisis de la información contenida en el data set con el propósito de entender su contenido , detectar patrones , valores , faltantes, anomalías en los datos y relacione s entre variables antes de construir los modelos , así como el posible desbalance de clases predictoras . Si bien es cierto la versión del dataset que contiene los datos procesados implicaría no tener la necesidad de validar datos faltantes o anomalías en los datos , es importantes que no saltemos esta fase asumiendo como verdad absoluta que el data set se encuentra sin errores, mas bien como buena practica vamos a ejecutar las fase del análisis exploratorio de datos sin saltarnos ningún paso.