Klasyfikacja nowotworów na podstawie mutacji somatycznych i ekspresji RNA

Anna Korczyńska

Zad 1. Przygotowanie danych

1. Filtracja genów onkogennych

Na podstawie filtracji genów z OncoKB uzyskałam ID 125 genów

2. Pobranie surowych mutacji

Wszystkich pobranych mutacji było 48912.

3. Format Parquet

4. Charakterystyka etykiet

Jak było spodziewane po treści zadania, różne klasy primary site mają bardzo różne liczebności. Najbardziej liczna Breast - 155358, najmniej np. Nasal cavity and middle ear - 126. Łącznie rekordów z danymi o ekspresji było 1448874.