Master Universitario in: "ANALISI DATI PER LA BUSINESS INTELLIGENCE E DATA SCIENCE" A.A. 2017-2018

Titolo della tesi: Valutazione prognostica nella malattia ematologica: Applicazione di strumenti di data mining su trial clinico nel mieloma multiplo

Autore: Salvatore Femiano

Abstract:

Nel campo della ricerca onco ematologica, la A.U.O. Città della Scienza e della Salute di Torino risulta essere uno dei poli più avanzati e prestigiosi del territorio nazionale. All'interno di esso troviamo il Centro Oncologico Ematologico Subalpino (COES), la cui missione è quello di assistere e seguire il paziente per tutto il percorso di cura. Parallelamente a tale attività vengono realizzati studi clinici atti a sperimentare l'efficacia di nuovi farmaci o a effettuare confronti tra nuove schemi chemioterapici e terapie di riferimento.

Presso il dipartimento di Biotecnologie Molecolari e Scienza per la Salute Divisione universitaria di Ematologia è in essere un studio sperimentale randomizzato dal nome "FORTE" in pazienti giovani affetti da mieloma multiplo in cui vengono messi a confronto l'induzione di carfilzomib-lenalidomide-desametasone (KRd) seguita dall'induzione di autotrapianto di cellule staminali ASCT e KRd con carfilzomib- ciclofosfamide-desametasone (KCd) seguito da consolidamento ASCT e KCd a loro volta raffrontati con il trattamento con KRd come induzione e carfilzomib più lenalidomide con lenalidomide come mantenimento.

Il project work oggetto della tesi è stato contestualizzato in questo studio sperimentale multicentrico. In particolare sono state estratte trentaquattro variabili alla base-line di cui quattro di outcome. Essendo il database dello studio non ottimizzato per data quality sono state necessarie attività di ETL e data cleaning.

A seguire si è proceduto con un analisi descrittiva delle singole variabili per studiarne le diverse distribuzioni ed un analisi bivariata con le variabili di outcome per determinare eventuali associazioni. Terminata questa fase si è deciso di scegliere una delle quattro variabile di outcome ritenuta più significativa e le relative variabili ad essa associata.

Dopo aver appurato che tra queste ultime non vi fossero correlazioni significative si è passati all'applicazione di algoritmi di machine learning. Sono stati impiegati k-nearest neighbors (knn) e decision tree(configurati per ottenere una la migliore accuratezza), per effettuare una classificazione supervisionata. Gli alberi decisionali si sono rivelati più accurati del Knn ed il modello ottenuto nella fase di addestramento è risultato abbastanza aderente alle regole epidemiologiche che definiscono la variabile di outcome scelta per l'analisi.