

Master Universitario in "ANALISI DATI PER LA BUSINESS INTELLIGENCE E LA DATA SCIENCE"

A.A. 2019/2020

Titolo della tesi: DATA DRIVEN RULE GENERATION – Utilizzo dell'algoritmo Apriori per l'identificazione di regole di Data Quality.

Autore: Alberto Ferrari

Abstract

Questo elaborato è stato scritto durante il tirocinio del candidato presso l'azienda Irion s.r.l., software vendor di un unico prodotto: Irion EDM.

In questa tesi, a introdurre il contesto, vi è una panoramica su un concetto emergente come la Data Fabric, una struttura per la gestione e l'integrazione di dati non soltanto strutturati.

All'interno di una Data Fabric i dati vengono organizzati, ma è necessario anche controllarli in maniera periodica affinché se ne possa trarre il massimo valore. La materia che si occupa di misurare la qualità dei dati è la Data Quality e spesso i processi di misurazione sono ripetitivi e richiedono molto tempo da parte degli utenti, per questo sempre più spesso le aziende sfruttano tecniche di Machine Learning e Artificial Intelligence per automatizzare e migliorare alcuni processi.

L'obiettivo del lavoro è valutare l'utilizzo dell'Association Rule Mining e in particolar modo dell'algoritmo Apriori per l'identificazione di regole di Data Quality all'interno di un dataset.

L'intero lavoro è stato svolto mediante l'ausilio di R. e del software Irion EDM.