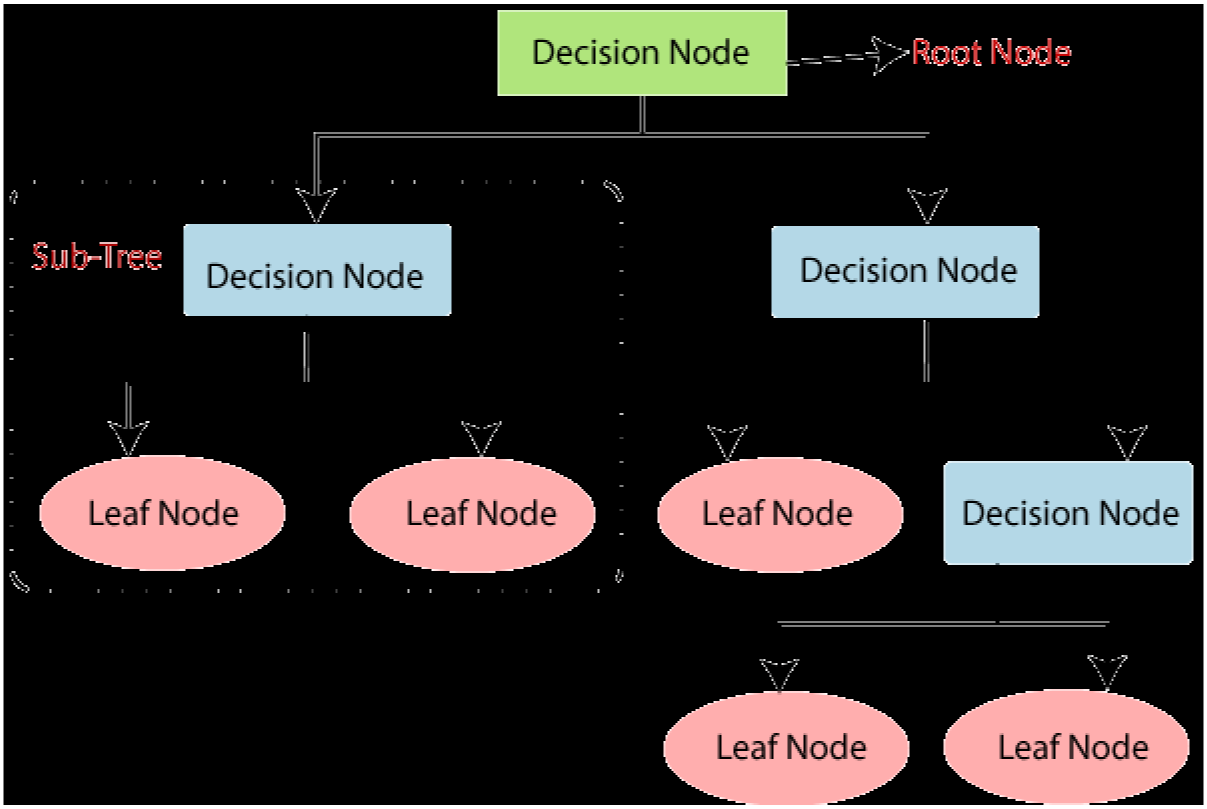
CSI4142 Phase 4 - Data Mining and Categorization - Group 19

Part A - Preprocessing - Documented by Simon

Through phase 2 and 3, we have done extensive data cleaning and normalizing of the tables. During Phase 2, we dropped rows that were a different range of dimensions than the merged dataset. This allowed to closely follow the dimensions GEO, REF-DATE and NAICS without having scrambled or incoherent data for the data observation. This allowed us to merge the datasets correctly, and saved us time when accessing the dataset for categorization. Since we have used official datasets from the government of canada that are often updated and cleaned, the datasets were extremely clean for mining purposes afterwards. We were able to find interesting visualizations of the data as employment and manufacturing sales are extremely comparative fact tables.

Part B - Data Mining Conclusions - Documented by Simon

Interesting “data nuggets” that were induced while running the data mining phase were plentiful. Although we had technical difficulties in the background through Google Colab, we were able to retrieve deferred algorithms, such as the Decision Tree, Gradient Boosting and Random Forest algorithms to process some visual information for the paper. Thanks to the DecisionTreeClassifier algorithm, we were able to predict the number of employers a certain NAICS sales of manufactured items would have, the accuracy depending on the province or year. Using the GradientBooster, the results had a weaker accuracy, and the Random Forest classifier had the slowest and least accurate results. The reason for this slow down is that the random forest uses no prior data for prediction, and uses a random categorizing for the outcome. GradientBooster uses rank boosting, whilst the sales’ spread and randomness rate makes the data at any classification level iffy even until the end. Decision tree seems to branch out the possibilities evenly and fairly and gives the result the fastest and most accurate for those reasons.



Thanks to these algorithms, we were able to classify the datamart's general information in an intuitive way that would be very beneficial for government officials and industry investors alike.