Matthew Naughton

**Portfolio: Searching for Similarity**

Decision trees are supervised and non-parametric, and have a structure similar to a flowchart. These trees are easily interpretable because you can see each “decision” the machine made and why it got what it got. kNN, or K-Nearest Neighbor, is also non-parametric and supervised. In kNN, the k is a constant defined by the user.

Both methods being non-parametric mean the distribution of data cannot be defined in a few parameters. (No assumptions) Decision trees are better are finding non-linear relationships, making them worse than say linear regression. In addition, bigger trees mean more overfitting, which means we must prune it. kNN on the other hand becomes slower with larger datasets, and becomes sensible to outliers.

kNN works with regression by approximating the relationship between independent variables and the continuous outcome via the average of the observations. Decision trees will regress the data by asserting true or false to specific questions, causing it to branch off – like a tree- into a true and false.