## **Classifiers**

**Damanpreet Singh**

Survival of patient’s study

The dataset contains cases from a study that was conducted between 1958 and 1970 at the University of Chicago's Billings Hospital on the survival of patients who had undergone surgery for breast cancer.

1. Age of patient at time of operation (numerical)
2. Patient's year of operation (year - 1900, numerical)
3. Number of positive axillary nodes detected (numerical)
4. Survival status (class attribute) 1 = the patient survived 5 years or longer 2 = the patient died within 5 years.

**Part 1: k-nn/ k nearest neighbour:**

The raw data initially included the Year the patient had operation; it wasn’t really meaning full so had to drop that column. 10 test runs were performed for each permutation of parameters. Both normalized and unnormalized data was used for test runs in this test. Testing and training data was split into 75/25 ratio and data was shuffled before split to randomize each set. The Accuracy was calculated after comparing labels collected after majority voting form testing and training data. If you care fully look at the permutation 6,7,8 they all used unnormalized data and that resulted in overfitting this could be most likely because data was mainly skewed to one side. (Function def unnormvisualizer(testset): can be used to plot the data and verify results, provided at the bottom of Assignment2A). The following table details all the findings.



**Part 2: Decision Tree**

The Year the patient had operation from this test as well as it wasn’t useful. Hype tuning was attempted with 8 different permutations as provided in the table below. Each tuning was tested with 5 runs each.



**Discussion**

**Part 1:** Test 3 where parameters were: votes = 5 using Makowski Distance on normalized data yielded good accuracy. When multiple tests were attempted, accuracy fluctuated between 100-90% which tells the parameters are really effective and avoid any biases. Normalized data must be used for this dataset to fix skewness.

Part 2: There were no clear winners while using Decision tree algorithm as all the test averaged around 70%. Using criterion = ‘entropy’, splitter =’best ’ yielded the best result with 71.79%. Test 3 of Knn classifier has been the best among both and its mainly because there were only two labels 1,2 and I believe need more to yield a good accuracy.

**Future Work**

Next, I would like to test how Decision tree look like using graph viz and would like to what were the decision nodes, to do a better tunning for future tests.