Andy Le - 000805099

Heart Failure Clinical Records Report

Data Description

We are using the Heart Failure Clinical Records Data Set. The data set features and their descriptions are as follows:

- Age: age of the patient (years)
- Anaemia: decrease of red blood cells or hemoglobin (boolean)
- High Blood Pressure: if the patient has hypertension (boolean)
- Creatinine Phosphokinase (CPK): level of the CPK enzyme in the blood (mcg/L)
- Diabetes: if the patient has diabetes (boolean)
- Ejection fraction: percentage of blood leaving the heart at each contraction (percentage)
- Platelets: platelets in the blood (kiloplatelets/mL)
- Sex: woman or man (binary)
- Serum Creatinine: level of serum creatinine in the blood (mg/dL)
- Serum Sodium: level of serum sodium in the blood (mEq/L)
- Smoking: if the patient smokes or not (boolean)
- Time: follow-up period (days)
- [target] Death Event: if the patient deceased during the follow-up period (boolean)

Tests/ Results

Part 1: k-nearest Neighbor

Different K and Euclidean Distance

I calculated the default K value of 13 by getting the square root of the data set size. And I also tested double the value of K at 27 and half of K at 7 to test the first batch. This test will be using the Euclidean Distance to calculate distance

K Value	AVG	Run 1	Run 2	Run 3	Run 4	Run 5
13	52.35	51.72	47.06	47.54	68.75	46.67
27	50.95	48.48	61.11	49.38	47.19	48.57
7	56.18	55.56	55.74	55.00	53.61	60.96

Different K and Manhattan Distance

This Test will be using the Manhattan Distance along with different K values

K Value	AVG	Run 1	Run 2	Run 3	Run 4	Run 5
13	60.47	61.80	56.52	100.00	55.17	48.85
27	50.86	52.00	50.36	45.16	50.53	56.25
7	54.35	50.43	58.25	44.83	57.14	61.11

Different K and Euclidean Distance and Normalized Data

This test will be using the Euclidean Distance along with different K values and Normalized Data

K Value	AVG	Run 1	Run 2	Run 3	Run 4	Run 5
13	61.34	57.78	60.40	57.94	71.43	51.13
27	58.55	60.00	53.85	61.11	59.10	58.70
7	46.68	25.00	50.00	52.63	57.14	46.65

Different K and Manhattan Distance and Normalized Data

This test will be using the Euclidean Distance along with different K values and Normalized Data

K Value	AVG	Run 1	Run 2	Run 3	Run 4	Run 5
13	56.80	53.49	56.41	61.53	56.96	55.56
27	49.28	56.18	47.61	47.06	58.07	37.50
7	60.82	57.69	66.67	56.86	56.20	66.67

Part 2: Decision Trees

Max_leaf_nodes = 2
Min_samples_leaf = 3
Min_impurity_decrease = 0.5

AVG	Run 1	Run 2	Run 3	Run 4	Run 5
45.00	49.05	42.50	39.02	45.10	43.48

Max_leaf_nodes = 2
Min_samples_leaf = 3
Min_impurity_decrease = 2.0

AVG	Run 1	Run 2	Run 3	Run 4	Run 5
45.10	44.64	46.15	43.64	36.84	49.02

Max_leaf_nodes = 2
Min_samples_leaf = 10
Min_impurity_decrease = 0.5

AVG	Run 1	Run 2	Run 3	Run 4	Run 5
44.60	34.10	47.62	46.88	46.67	51.11

Max_leaf_nodes = 2
Min_samples_leaf = 10
Min_impurity_decrease = 2.0

AVG	Run 1	Run 2	Run 3	Run 4	Run 5
44.61	36.17	43.90	42.42	38.24	48.89

Max_leaf_nodes = 10
Min_samples_leaf = 3
Min_impurity_decrease = 0.5

AVG	Run 1	Run 2	Run 3	Run 4	Run 5
44.40	47.50	41.51	50.00	50.00	48.65

Max_leaf_nodes = 10
Min_samples_leaf = 3
Min_impurity_decrease = 2.0

AVG	Run 1	Run 2	Run 3	Run 4	Run 5
45.13	41.03	47.50	48.15	46.88	47.73

Max_leaf_nodes = 10
Min_samples_leaf = 10
Min_impurity_decrease = 0.5

AVG	Run 1	Run 2	Run 3	Run 4	Run 5
45.76	50.00	36.59	41.18	43.48	43.75

Max_leaf_nodes = 10
Min_samples_leaf = 10
Min_impurity_decrease = 2.0

AVG	Run 1	Run 2	Run 3	Run 4	Run 5
45.33	51.06	48.15	48.72	51.22	44.44

Discussion

The clear winner was the K = 13, Euclidean Distance with Normalized Data. While the clear loser was K = 7, Euclidean Distance with Normalized Data. Some

versions may be better than the others because of the normalization of data making numbers more comparable. Sometimes, normalizing made the test better while others made the test worse. It seems that normalizing the data with this data set doesn't clearly help the tests as both normalized and un normalized data have similar comparable results. When looking at K values, having K =7 to K = 13 (square root of data set size) seems to result in the best tests. There is never a time where K = 27 has the best average results. When comparing distances, Manhattan seems to perform a bit better compared to Euclidean only on normalized data sets, whereas Euclidean does better on non-normalized data sets. I would personally recommend K = 13, Euclidean Distance with Normalized Data as it had the best results. Overall K = 13 has seemed to be the best K value in all test cases on average as it has never been the lowest of the 3 K values. The Euclidean Distance seems to also be recommended as none of the average percentages dip below 50%.

There are no clear winners or losers for Decision Trees. The decision tree performed worse than the Knn tests. Some versions of the decision trees with max_leaf_nodes = 10 were better because limiting the number of leaf nodes provided better results. The other changes in properties doesn't seem to change the results of the test much at all. When comparing the decision trees and the kNN, the kNN performs much better and it seems to be because kNN is testing over similar examples and learning from the training instances. kNN is much slower and more expensive but can be well-tuned resulting in better results. In terms of the decision tree, I would recommend

Max_leaf_nodes = 10
Min_samples_leaf = 3
Min_impurity_decrease = 2.0

as it provided the best results. My final recommendation is that you should stick with the kNN recommendation because it has overall better results.

Future Work

If I had more time, I would want to explore a larger data set of this data. I feel like there was not enough data to have meaningful comparisons and tests. In terms of the kNN algorithm, I would play more around with the K value as it semed to have the most impact with my data set. And with the Decision Tree, I would've wanted to try out other modifiers and even more different versions to get a better understanding of how they worked and see if I could get better results.