Toby Towler Machine Learning Part 3 Report

Introduction

This report is a summary of my findings while evaluating different attribute values for various types of data classifiers. Testing was done using the MNIST dataset provided by the sklearn python library. The data was split, using 85% for 5 fold cross validation training with the remaining 15% used for testing the best set of attributes to determine a best case score for each classifier

The dataset

The dataset used in this report was a slimmed down version of the MNIST data provided by sklearn python library. The dataset consists of a series of 64 pixel grids representing number images from 0-9 for a total of 10 classes. Each pixel has its own attribute that can take 17 values, 0 to 16 depending on the "intensity" of the pixel. This means how shade of the pixel from black to white. There area total of 1797 images in this dataset

Implementation and experiments

I chose 3 classifiers from the sklearn library for this report, each with varying parameters.

RandomForest classifier

First I implemented the randomForest classifier. The parameters I changed for this classifier and their effects are as follows:

Number of estimators – the number of trees in the forest, higher is better

Maximum depth – the maximum number of nodes a branch may reach, effectively its length, higher is better

Maximum leaf nodes – the amount of leaves a tree can form, contributing how specific the tree can be, higher is better

Minimum samples to split – the population a node must have to be able to split into several nodes, lower is better

K nearest neighbour classifier

Secondly i used the K nearest neighbour classifier. The parameters I changed for this classifier and their effects are as follows:

Number of Neighbours – number of neighbours that can be used to query data, higher is better

Leaf size – size of the leaves passed to tree algorithms, higher or lower could be better depending on the nature of the problem

P – power parameter of the Minkowski metric, higher or lower could be better, depending on the nature of the problem

Decision tree classifier

Lastly, the decision tree classifier. The parameters I changed for this classifier and their effects are as follows:

Maximum depth – the maximum amount of levels a tree can have, higher is better

Minimum samples per split – the minimum number of samples needed for a node to branch into 2 lower nodes, lower is better

Maximum leaf nodes – maximum number of nodes a tree may have at one time, higher is better

Minimum samples per leaf – number of cases a node must have in both child nodes in order to split itself

For all classifiers that needed it, a random state was kept constant for the sake of replicability. I used 3 parameter combinations for each classifier, the results for each are shown below. For each set of parameters, the classifiers were trained on a combination of subsets and tested on an excluded subset. This happed 5 times and the mean of these scores are the accuracies used to determine which combination was the most successful

RandomForest

Estimators	Depth	Max leaf nodes	Min sample split	Random state	accuracy
50	5	5	5	26	0.7986
150	12	12	12	26	0.8798
300	20	20	20	26	0.8993

Since the the last input set score the highest accuracy, it was retrained with the whole data set and scored 0.9333

KNN

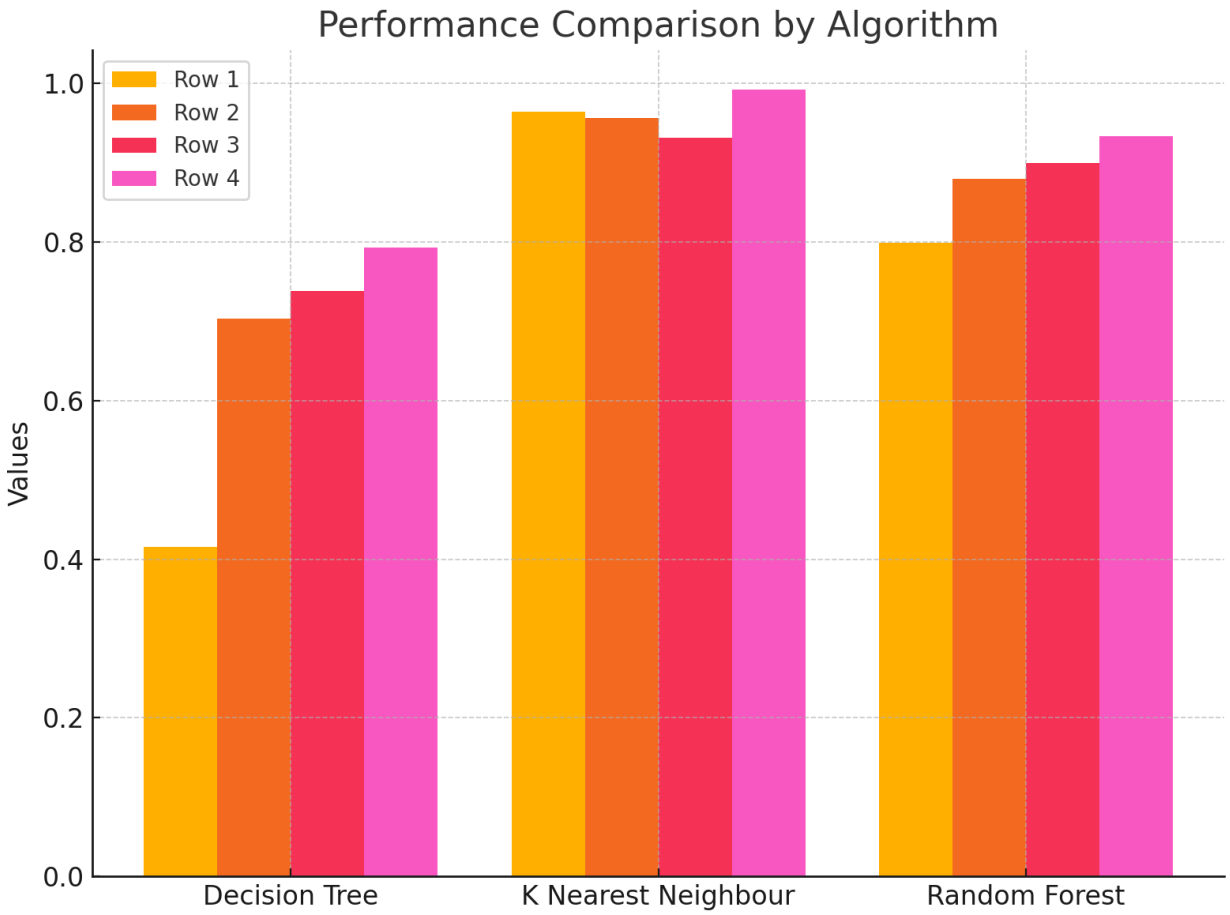
neighbours	Leaf size	p	accuracy
5	30	2	0.9644
15	50	5	0.9560
30	100	12	0.9316

The first input was the highest scoring, when tested with the full data set it achieved an accuracy of 0.9926

Decision tree

Max depth	Min sampe split	Max leaf nodes	Min sample leaf	Random state	accuracy
5	2	5	5	26	0.4152
12	4	12	12	26	0.7034
25	8	25	20	26	0.7380

The bottom row got the highest accuracy score, when retrained with the full data set it scored an accuracy of 0.7926.



Conclusion

From the data collected over the span of making this report, I can conclude that for classifying encoded image data distributed across several classes is the k nearest neighbour algorithm. It achieved a higher score in all 3 parameter sets than any other classifier by a large margin. Scoring a very impressive 99.3% across the entire dataset. Each classifier was tested on several data inputs on all parameters and average to give a fair and consistent end result.