Data Mining & Machine Learning F20DL

Group 4
Lewis Wilson, Sam Fay-Hunt, Kamil Szymczak, Chun Man

November 30, 2020

Contents

1 Variation in performance with size of the training and testing sets

2 Variation in performance with change in the learning paradigm (Decision Trees versus Neural Nets)

3 Variation in performance with varying learning parameters in Decision Trees

3.1 J48

3.2 Random Forest

"max_features" has the options "auto", "sqrt" and "log2". From (Figure 1) we can see that "sqrt" has a higher accuracy overall, the accuracy of "log2" varies between the lower end and the median accuracy value.

"min_samples_split" has very little impact on accuracy.

"criterion" gives a perfect negative correlation with respect to accuracy. Correlation values - [Gini = -0.404], [Entropy = 0.404].

"n_estimators" which defines the number of trees in the forest seems to have very little correlation but high importance.

"min_samples_leaf" gives a strong negative correlation in terms of accuracy, meaning the higher minimum samples at a leaf node, the lower the accuracy.

"min_weight_fraction_leaf" has a somewhat positive correlation.

4 Variation in performance with varying learning parameters in Neural Networks

4.1 Linear Classifier

4.2 Multilayer Perceptron

5 Variation in performance according to different metrics (TP Rate, FP Rate, Precision, Recall, F Measure, ROC Area)

Appendices

A Appendix A

A.1 Workload split

Team member	**Involvement**
Lewis Wilson	text here
Chun Man	text here
Sam Fay-Hunt	text here
Kamil Szymczak	text here

As a team we are happy with everyone's contributions to the project. All team members were punctual and showed up to all scheduled meetings. Sam took the lead as project manager throughout the project delegating the workload and providing support to others.

A.2 Random Forest Parameter Importance

Parameter Config	Importance	Correlation
min_samples_split	0.05	0.306
min_samples_leaf	0.375	-0.725
n_estimators	0.015	0.092
min_weight_fraction_leaf	0.014	0.123

Parameter Config	Importance	Correlation
max_features.value_sqrt	0.010	0.563
max_features.value_log2	0.959	-0.752
criterion.value_entropy	0.456	0.404
criterion.value_gini	0.544	-0.404

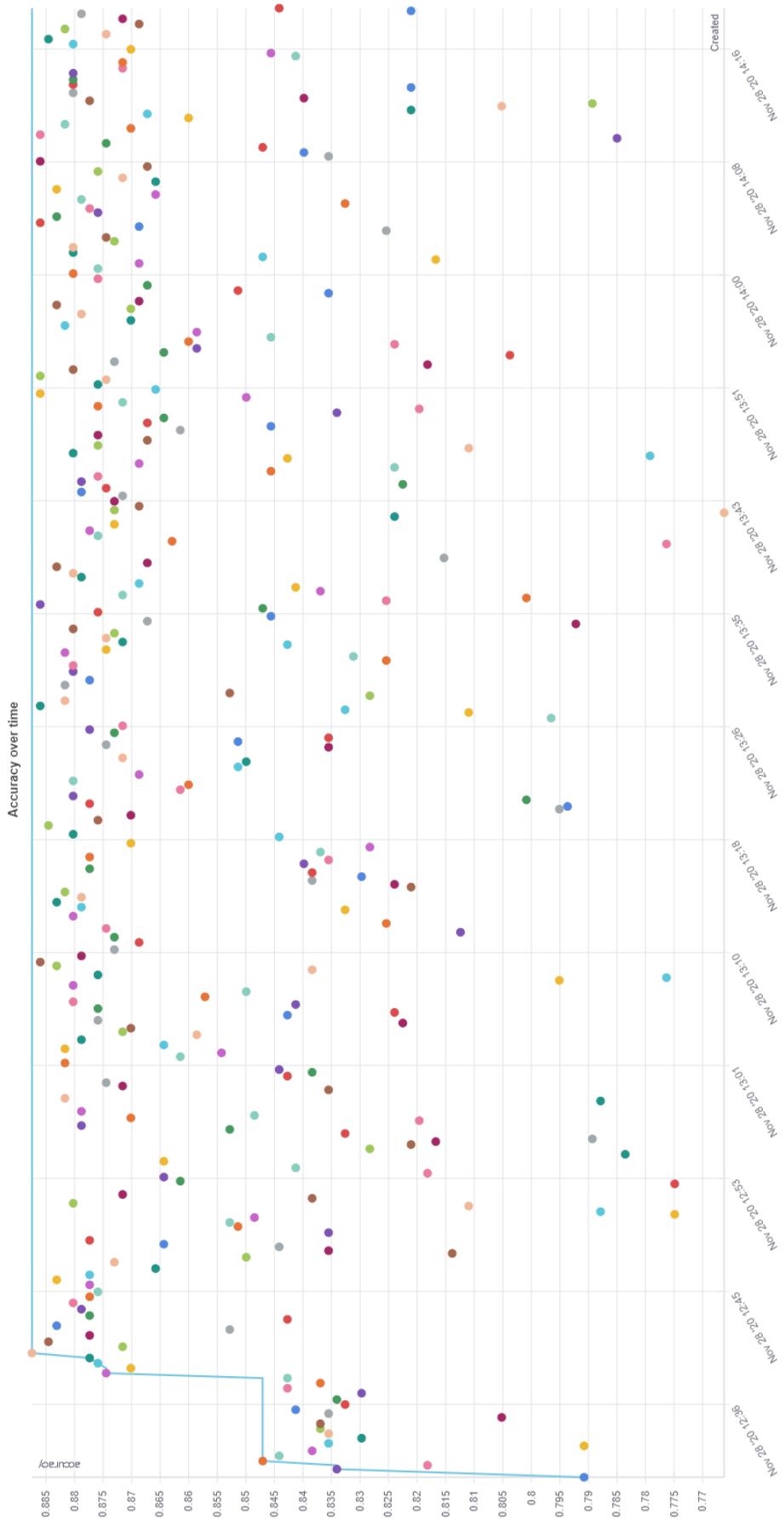

Figure 1: Accuracy over time - Random Forest

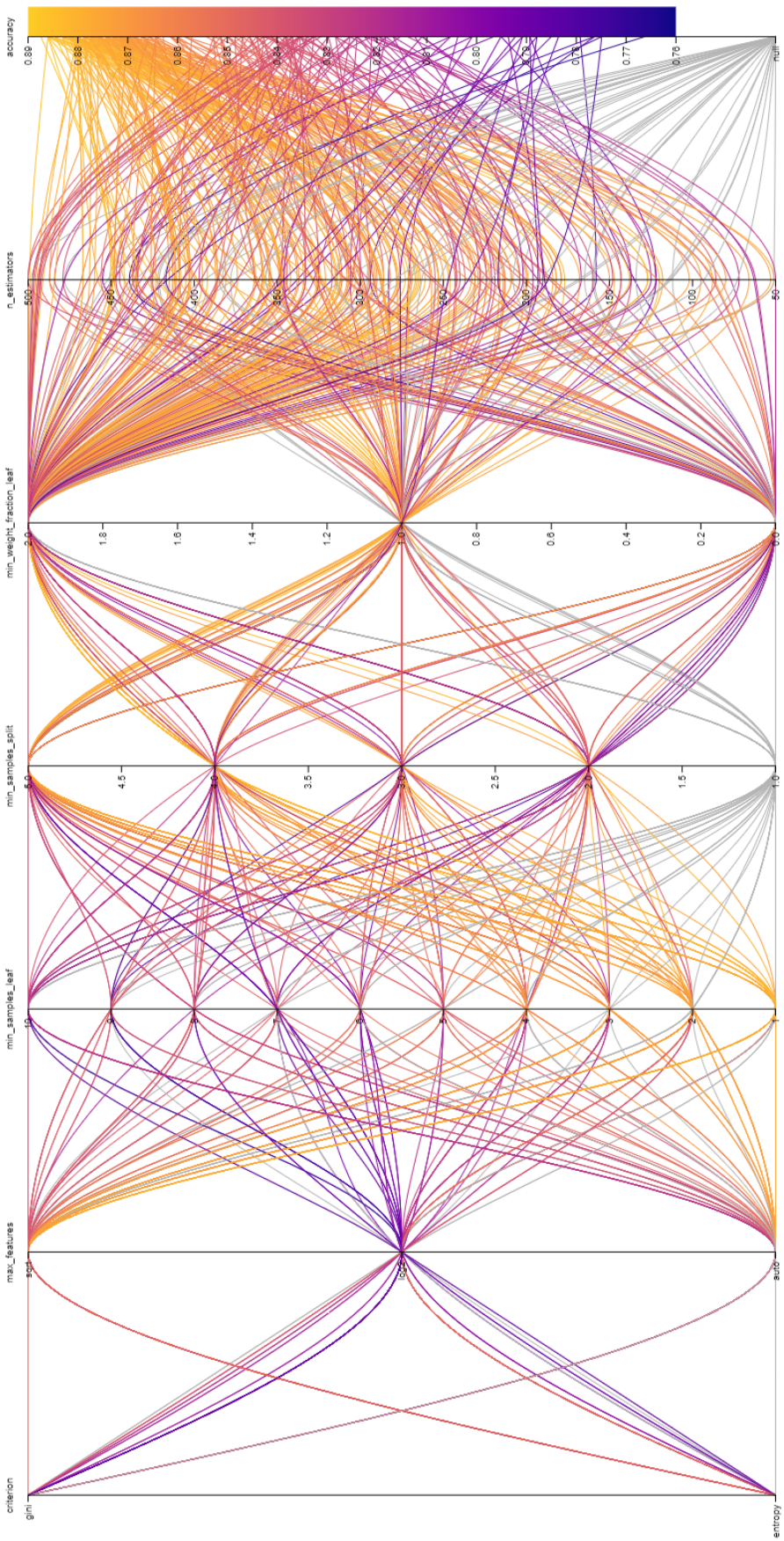

Figure 2: Random Forest Parameters