Data Mining & Machine Learning F20DL

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

December 2, 2020

Contents

1 Introduction

We used Weights and Biases (`https://wandb.ai/home`) to run these experiments which allowed us to generate the graphs shown in this document.

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

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

4 Variation in performance with varying learning parameters in Decision Trees

4.1 J48

4.2 Random Forest

Parameter	Conclusion
max_features	Contains 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	The minimum samples required to split a node has very little impact on accuracy.
criterion	Gives a perfect negative correlation with respect to accuracy. Correlation values being [Gini = -0.404], [Entropy = 0.404].
n_estimators	This is defined as the number of trees in the forest, it 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 to accuracy. e.g. total weight required at a leaf node varies between 76% and 89% accuracy

See Parameter Importance (Figure C.1)

5 Variation in performance with varying learning parameters in Neural Networks

5.1 Linear Classifier

5.2 Multilayer Perceptron

Parameter	**Conclusion**
alpha	This has a positive correlation to accuracy as higher alpha value equates to higher accuracy.
solver	lbfgs is the most accurate value of this parameter with a strong positive correlation out of the three (lbfgs, adam, sgd).
max_iter	The maximum number of iterations - In general, higher accuracy can be achieved with a larger amount of max iterations.
activation	Out of the four activation functions (relu, tanh, identity and logistic), relu is the only one with a positive correlation, giving the highest accuracy overall.
learning_rate	'adaptive' achieves the highest accuracy while, 'constant' and 'invscaling' vary widely.
hidden_layer_sizes	Has a negative correlation - the number neurons in the n-th hidden layer has no effect on accuracy.

6 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.

B J48

C Random Forest

C.1 Random Forest Parameter Importance

Parameter Config	Importance	Correlation
min_samples_split	0.005	0.306
min_samples_leaf	0.375	-0.725
n_estimators	0.016	0.092
min_weight_fraction_leaf	0.013	0.123

Parameter Config	Importance	Correlation
max_features.value_sqrt	0.001	0.563
max_features.value_log2	0.565	-0.752
criterion.value_entropy	0.012	0.404
criterion.value_gini	0.012	-0.404

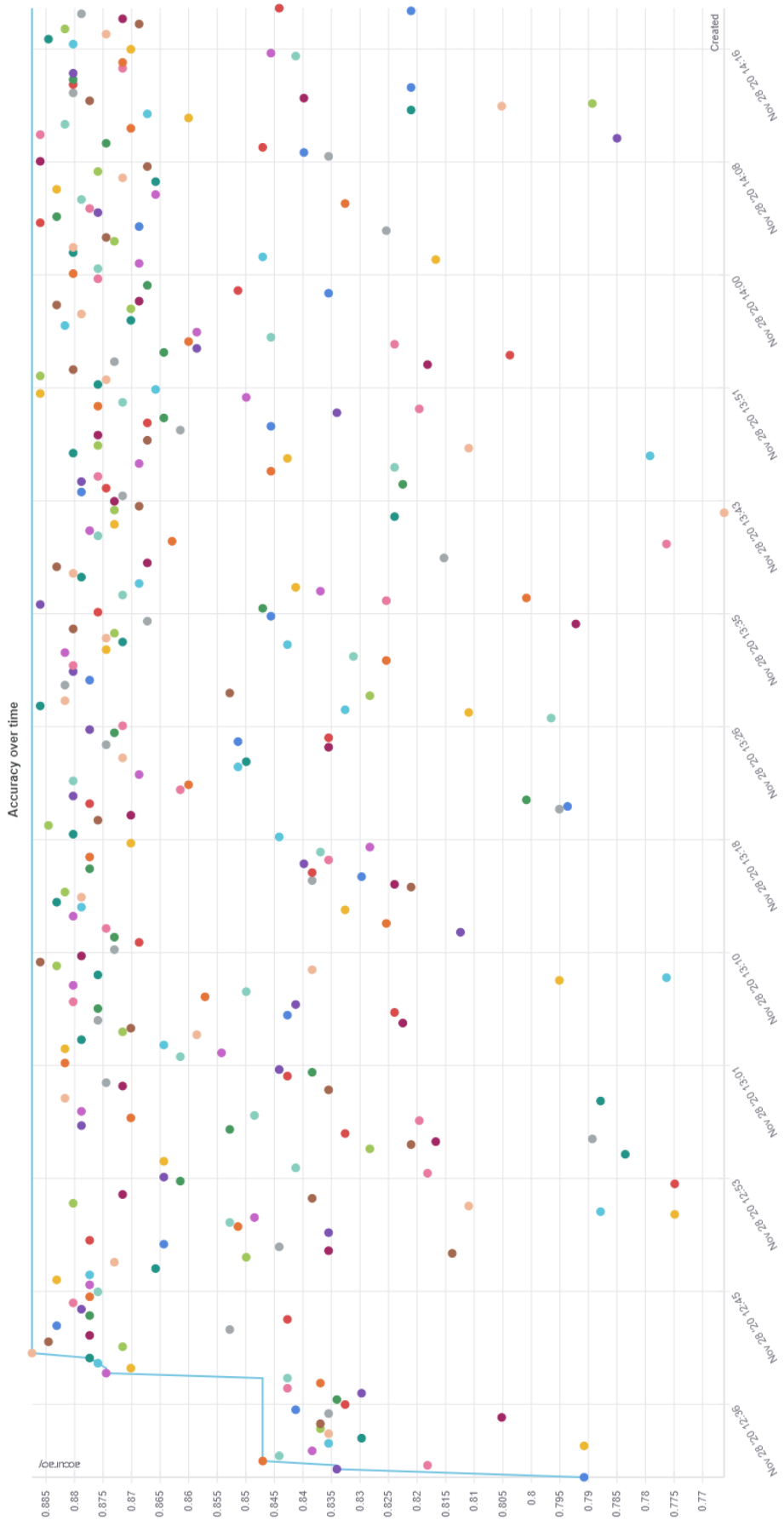

Figure 1: Accuracy over time - Random Forest

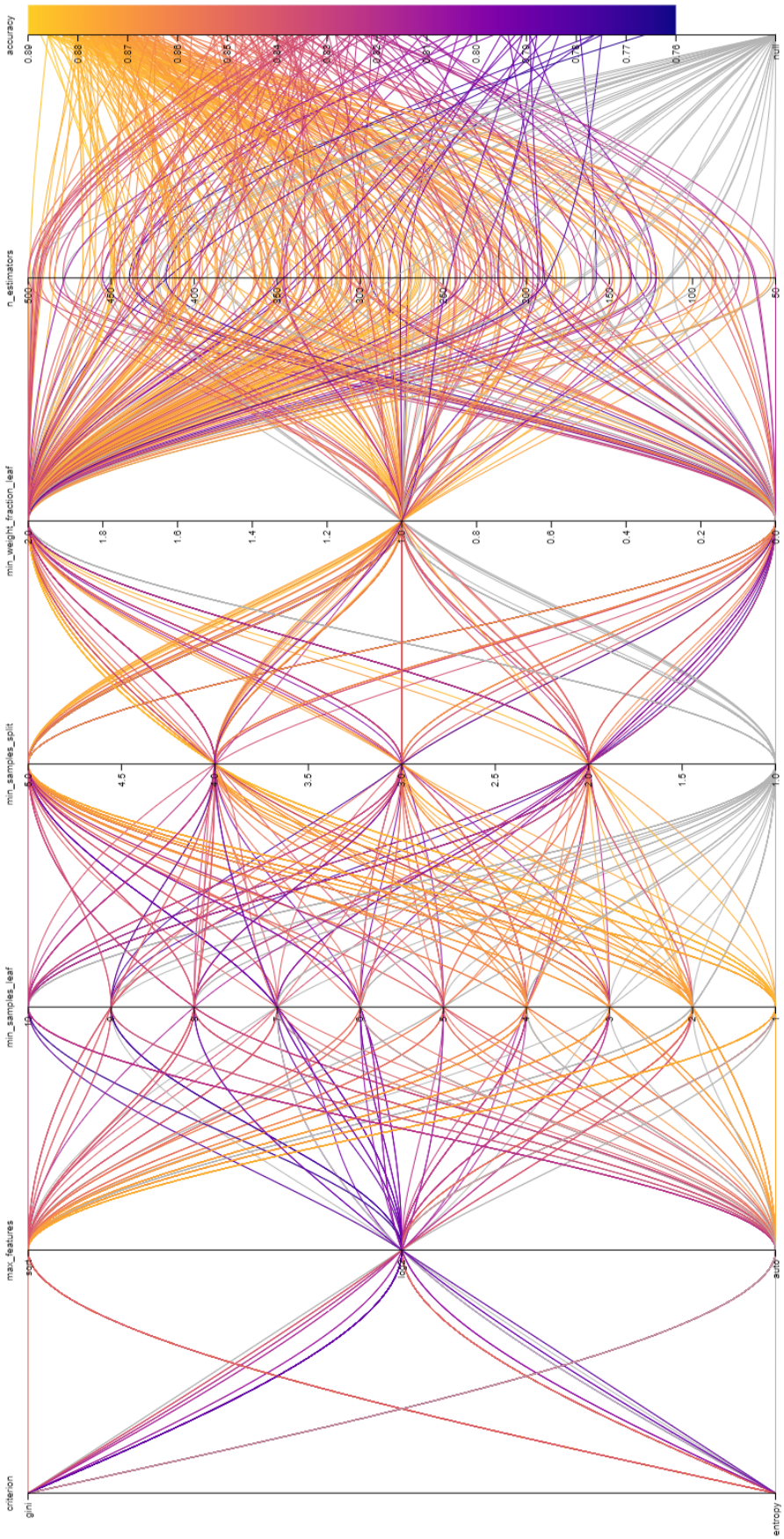

Figure 2: Random Forest Parameters

D Linear Classifier

E Multilayer Perceptron

E.1 Multilayer Perceptron Parameter Importance

Parameter Config	Importance	Correlation
hidden_layer_sizes	0.095	-0.101
max_iter	0.091	0.072
alpha	0.061	0.171

Parameter Config	Importance	Correlation
solver.value_lbfgs	0.530	0.728
solver.value_adam	0.027	-0.245
solver.value_sgd	0.024	-0.640
activation.value_identity	0.098	-0.169
activation.value_relu	0.033	0.301
activation.value_tanh	0.018	-0.035
activation.value_logistic	0.006	-0.270
learning_rate.value_adaptive	0.012	0.507
learning_rate.value_constant	0.004	-0.442
learning_rate.value_invscaling	0.001	-0.224

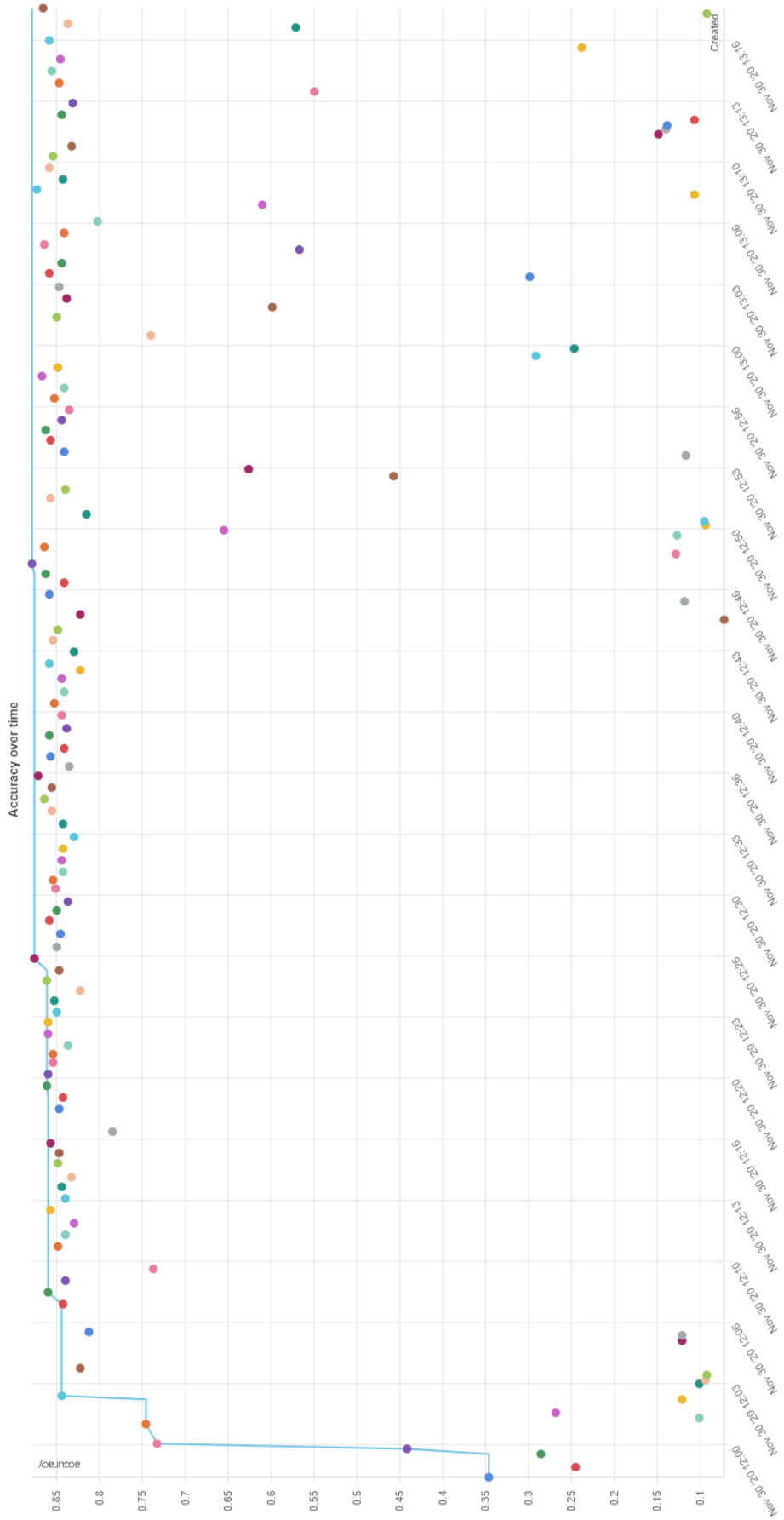

Figure 3: Accuracy over time - Multilayer Perceptron

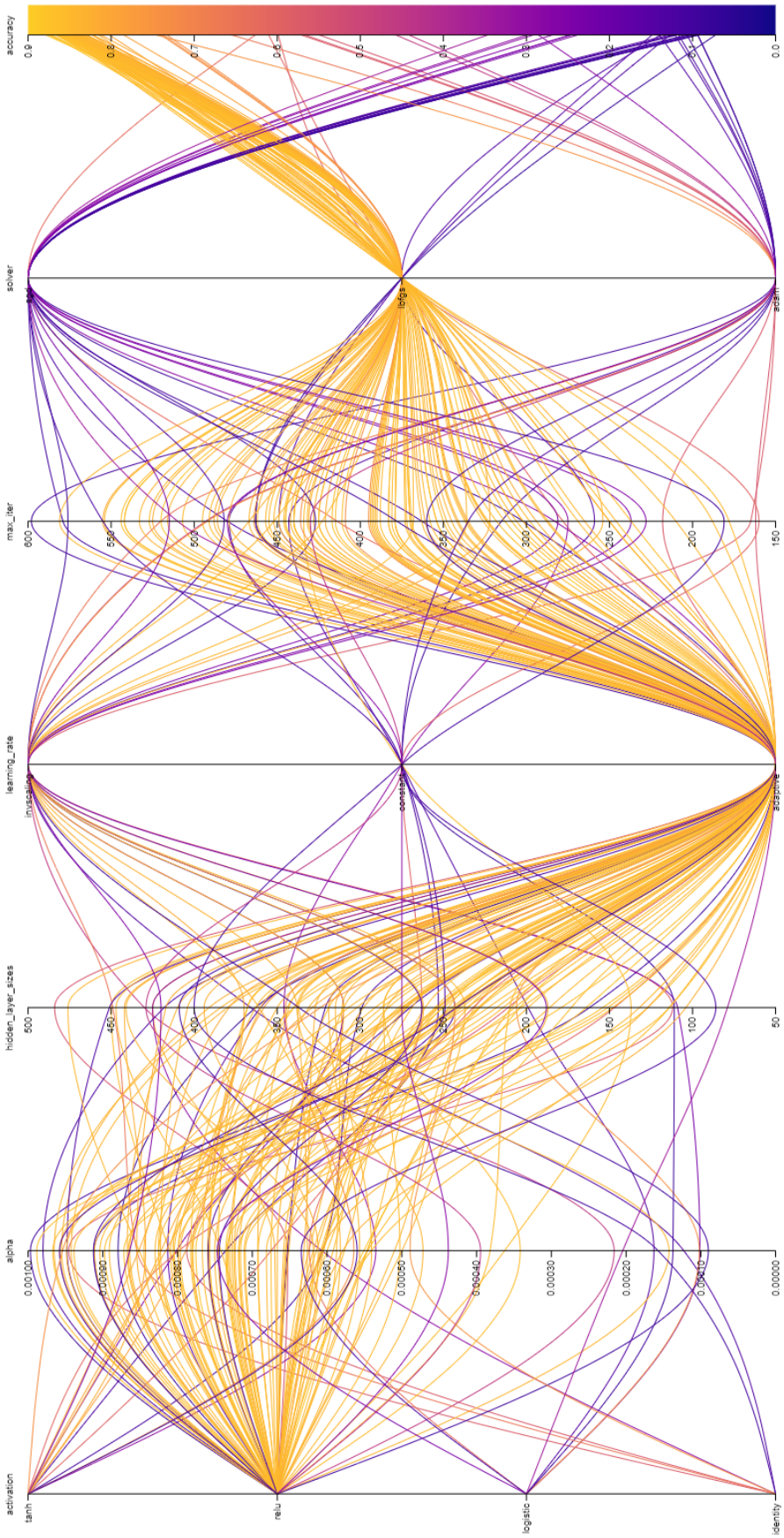

Figure 4: Multilayer Perceptron Parameters