Judith Grieves

Predicting Student Grades – a Comparison of Decision Trees and Naïve Bayes

Supplementary Material

Glossary

Term	**Description**
Attribute	An item of data with a label and a defined set of values. Also referred to as a column (database) or feature.
Binary Classifier	A classifier with two possible values, usually 0/1 or yes/no.
Categorial	A class of variables whose valid values are a finite set of discrete values.
Classifier	A data attribute which labels a data instance with one of a discrete set of values.
Conditional Probability	The probability of a given event, dependent on a separate event. The probability of event A, given that event B has occurred. Written as P(A|B).
Confusion Matrix	A square matrix consisting of the TP, TN, FP, FN counts from training or testing.
Correlation	A statistical relationship between two variables, describing the extent to which the increase in value of one is related to the increase or decrease in the value of the other.
Cross-Validation Error	The errors obtained for each fold in K-fold cross validation.
Disjoint	In data, disjoint describes sets of values that do not overlap.
FN (False Negative)	A set of binary classification results where the algorithm's predicted value is negative but the known value is positive. The prediction is therefore incorrect.
FN (False Positive)	A set of binary classification results where the algorithm's predicted value is positive but the known value is negative. The prediction is therefore incorrect.
Generalisation	A desirable feature of a machine learning model whereby the trained model has accurately modelled the underlying rules of the problem and is able to obtain correct predictions on new, previously unseen data.
Hyperparameters	Variable features of machine learning models which can be specified in training and used to evaluate the optimum features of a model for the problem at hand.
Information Gain	In Decision Trees, information gain is a means of assessing data attributes for their usefulness in classifying or dividing the data. There are multiple algorithms used to measure information gain.
Instance	A set of data features or attributes defining one set of results or thing. Also referred to as a row or record.
Kernel Smoothing	
K-Fold Cross Validation Error	A means of machine learning, such that the training data is divided into k subsets and the learning algorithm executed k times. Each execution trains using k-1 sets and validates results on the kth set. Validation statistics produced are a combination of the k result sets.

minLeafSize	MATLAB fitctree hyperparameter to set the minimum number of observations on each leaf of the resulting Decision Tree. Default value is 1.
minParentSize	MATLAB fitctree hyperparameter to set the minimum number of observations at each branch node in the resulting Decision Tree. Default value is 10.
maxNumSplits	MATLAB fitctree hyperparameter to set the maximum number of node splits in the resulting Decision Tree. Default value is the training dataset size – 1.
Over-Fitting	Over-fitting can be a feature of the training of a machine learning model when the learned algorithm takes too much account of noise in the data. This fails to produce a model that will generalise well, ie. make correct predictions for new, unseen data.
Post-Pruning	A method of optimising a Decision Tree after its creation by removing leaves and branches and reducing complexity. Usually carried out to reduce over-fitting and aid generalisation.
Predictive Variables	Attributes of a dataset used in a machine learning algorithm to predict the value of another attribute, the classifier.
Prediction Accuracy	The accuracy of a machine learning prediction is defined here as the proportion of correct predictions divided by the total number of predictions.
Sensitivity	In prediction accuracy, the sensitivity is the proportion of positive results that are correctly predicted, eg. For the Student experiments, the proportion of correct 'pass' predictions. Calculated by TP/(TP+FP).
Specificity	In prediction accuracy, the specificity is the proportion of negative results that are correctly predicted, eg. For the Student experiments, the proportion of correct 'fail' predictions. Calculated by TN/(TN+FN).
Supervised Learning	A method of machine learning where the labelling of instance is pre-defined and already known for a set of training data. The known labels can be used as a means of evaluating the predicted values.
TN (True Negative)	A class of binary classification results where the known value is negative and the algorithm's predicted value is correctly also negative.
TP (True Positive)	A class of binary classification results where the known value is positive and the algorithm's predicted value is correctly also positive.

Intermediate Results

Experimental Results by test/train split Percentage

	Decision Tree			Naive Bayes		
Test %	Average of Train Accuracy	Average of Test Accuracy	Average of Comb Accuracy	Average of Train Accuracy	Average of Test Accuracy	Average of Comb Accuracy
30	94.43	88.32	91.37	88.13	88.49	88.31
40	94.62	88.93	91.77	88.89	87.77	88.33
50	95.49	88.58	92.03	89.85	87.86	88.85

Fig. 1 Grid search results of varying training and testing split percentages

In running the experiments to evaluate the train/test split percentage, we see (Fig. 1) that the Decision Tree (DT) shows a small improvement for test accuracy for 60/40 but better training accuracy for 50% holdout. The Naïve Bayes (NB) shows more ambiguous results; the test accuracy is highest for 70/30, but the best training accuracy is for 50/50 split. Using the combined accuracy as a guide, for the best model, we choose 50/50 split.

Note that NB has a better balance between train and test accuracy and DT shows more overfitting across all values. We noted also in the program output, that the distribution of the classifier across the training and testing datasets is satisfactory for all the parameters tried (~85% pass).

Decision Tree Predictive Feature Selection

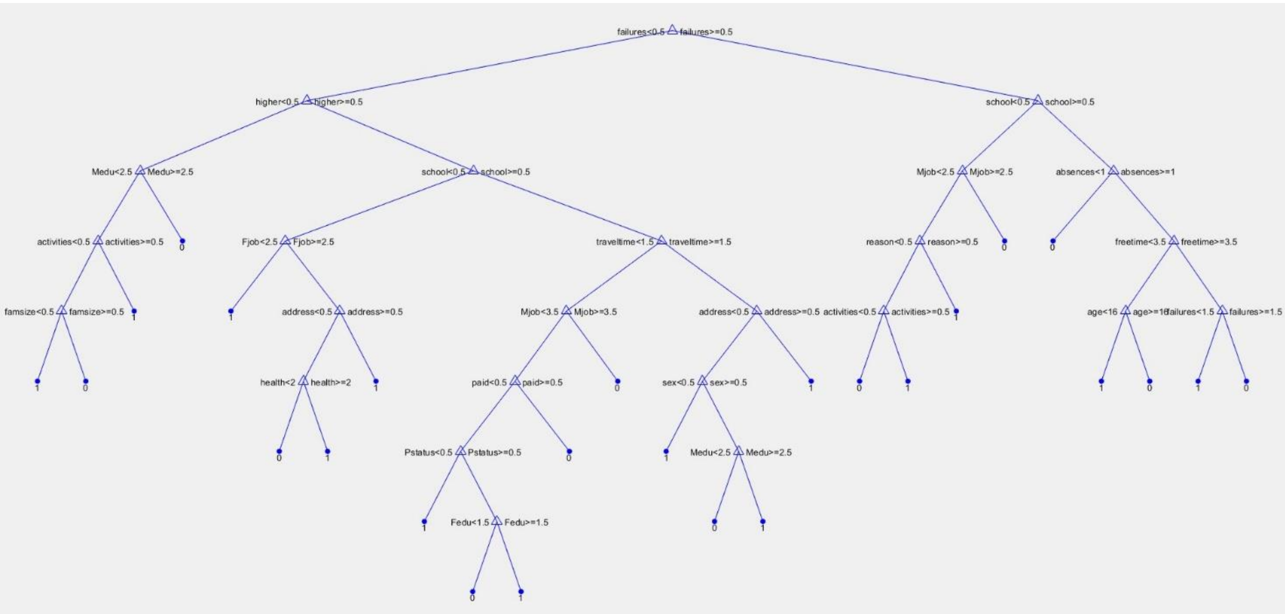
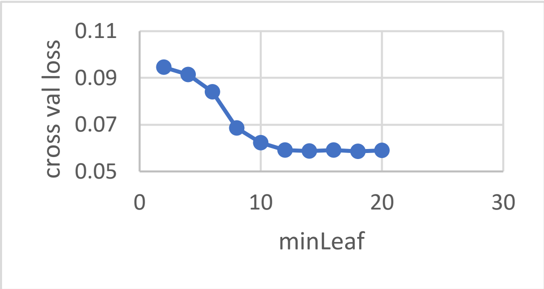
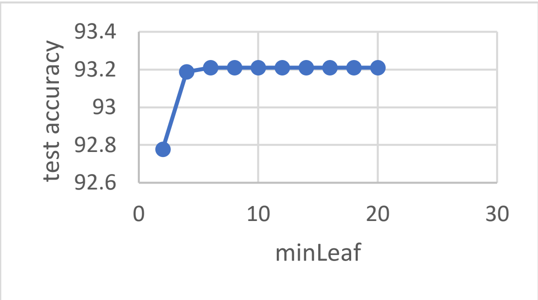


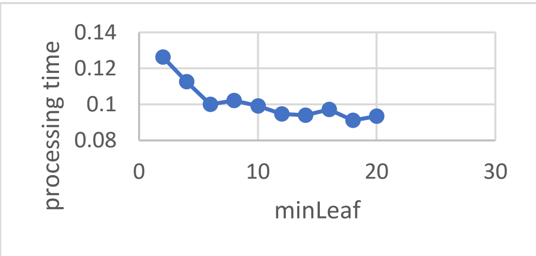
Fig. 2 Decision Tree diagram: Student data using all predictive variables but excluding previous grades, G1 and G2.

The DT tree diagram (Fig. 2) show us which predictive features have been used in the absence of the previous grades – this choice matches well with the variables most correlated with the final grade in our data analysis. The DT's ability to effectively pick the most predictive variables would be useful where there are many and could save extensive initial data analysis in further experiments.

Decision Tree Hyperparameter Grid Search – Performance Measures







To optimise the Decision Tree modelling, grid searches were carried out for the MATLAB DT hyperparameters minLeafSize, maxNumSplits and minParentSize. These parameters control the algorithm and hence the structure of the resulting tree.

Results for each parameter set were evaluated based on (a) minimising the cross validation error of the k-fold training algorithm (b) minimising the training time taken and (c) maximising the resulting Test Set prediction accuracy.

Fig. 3 shows an example of the analysis for minLeaf parameter values. In this case, minLeafSize=18 (default value = 1) gave optimal results across all benchmarks.

Similarly values maxNumSplilts=260 (default value = size of training set = 325) and minParentSize=20 (default = 10) were set for the training of the best models, to feed into

the comparison of machine learning methods.

Fig. 3 Plots of grid search experiment results for MATLAB minLeafSize

Decision Tree Hyperparameter Grid Search – Tree Structure

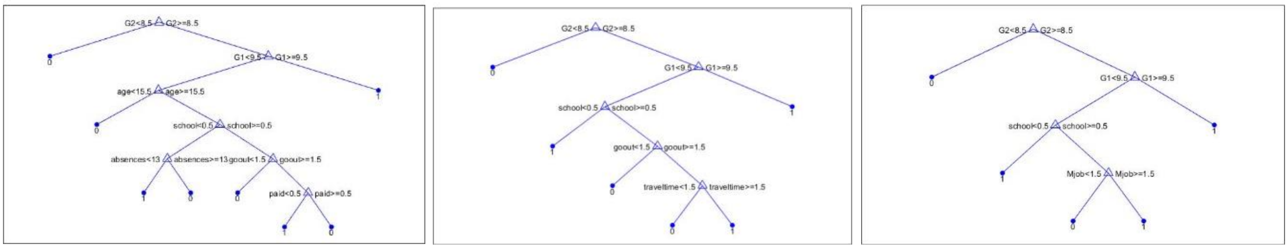


Fig. 4 Tree structure experiment results for MATLAB minLeafSize parameter values: (a) 2 (b) 4 (c) 6

Varying the DT hyperparameters (when using all predictive variables) also impacts the final tree structure and its number of nodes and splits. For example, Fig. 4 shows the results of varying the minLeafSize parameter value. In these examples, as the value increases, the tree structure simplifies with fewer nodes. This corresponds with the performance measure results above (Fig.3).