Assessment 3: Project

David Cole

UNSW
Sydney, Australia
z5294919

***Abstract*—This paper presents several tree-based classification models for determining the age of Abalone given other properties like weight, height, and sex.**

I. Introduction

The traditional method of determining the age of Abalone is a time consuming and destructive task, requiring the shell to be cut and stained to reveal a countable set of rings [1]. The number of rings is equivalent to the number of years of life. The models in this paper attempt to broadly classify an age bracket for specimens using readily available properties of the Abalone.

II. Methodology

A. Supervised Learning Task

The task is to classify an Abalone's age by other recorded features. In the original data set, age is an integer number of years of life. For this model, age has been grouped as follows to assist with the task of classification:

- Class 1: 0-7 years
- Class 2: 8-10 years
- Class 3: 11-15 years
- Class 4: >15 years

B. Data

Models use the University of California repository of Abalone data [1] as labeled examples to train the classifiers. The 'sex' feature was converted from categorical to numerical data to work with Sci-Kit Learn's tree classification model. All features were used for the task. An initial visualisation revealed a heavy class imbalance in the target feature, age: there are far fewer Abalone recorded in the youngest and eldest classes.

C. Performance

Both accuracy and F1 scores were used to validate the models as described in this paper.

D. Model Construction

Sci-Kit Learn's Decision Tree Classifier and Random Forest models were used to perform classifications. The default Gini criterion was used across both sets of models, with the random seed set initially to a constant for reproducibility and to compare across models. A simple classification tree model is first presented, followed by a random forest, and then a hyper-parameter grid search for each model to discover the most optimal learner. Sci-Kit Learn's balanced class weight feature was used to adjust for the original class imbalance in the age feature.

III. Results

A. Data Analysis and Modification

Table I displays a sample of the data in its original format.

As previously described, the 'sex' feature was changed from categorical to numerical data, and the original 'age' integers were placed into four classes. Table II outlines the number of samples in each age class. It is evident in this table that the data is heavily skewed towards the middle two age brackets, with few in the youngest and eldest classes.

A visualisation of the covariances of the transformed data is available in Fig. 1.

B. CART Classification Model

A single CART classification model was created to set a baseline for performance. The model used the balanced weight class feature to account for the age class imbalance, and had the following parameters:

```
{'ccp_alpha': 0.0,
'class_weight': 'balanced',
'criterion': 'gini',
'max_depth': 5,
'max_features': None,
'max_leaf_nodes': None,
'min_impurity_decrease': 0.0,
'min_samples_leaf': 100,
'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0,
'random_state': RandomState(MT19937)
at 0x2BB4B5340,
'splitter': 'best'}
```

The model has average performance, with a training accuracy of 54.31% and test accuracy of 54.83%. A representation of the tree model can be seen in Fig. 2.

A test accuracy higher than training accuracy suggests that improvement may be possible via hyper-parameter tuning. A grid search was undertaken to discover a more optimal model by 'early stopping' the trees with the 'minimum samples' and 'max depth' hyper-parameters. The best model from the search had the following parameters:

```
DecisionTreeClassifier(
class_weight='balanced', max_depth=10,
min_samples_leaf=10,
random_state=RandomState(MT19937) at 0x2BD6FF340)
```

TABLE I
SAMPLE OF ORIGINAL DATA

Sex	Length	Diameter	Height	Whole Weight	Shucked Weight	Viscera Weight	Shell Weight	Rings (Age)
M	0.455	0.365	0.095	0.5140	0.2245	0.1010	0.150	15
M	0.350	0.265	0.090	0.2255	0.0995	0.0485	0.070	7
F	0.530	0.420	0.135	0.6770	0.2565	0.1415	0.210	9
M	0.440	0.365	0.125	0.5160	0.2155	0.1140	0.155	10
I	0.330	0.255	0.080	0.2050	0.0895	0.0395	0.055	7

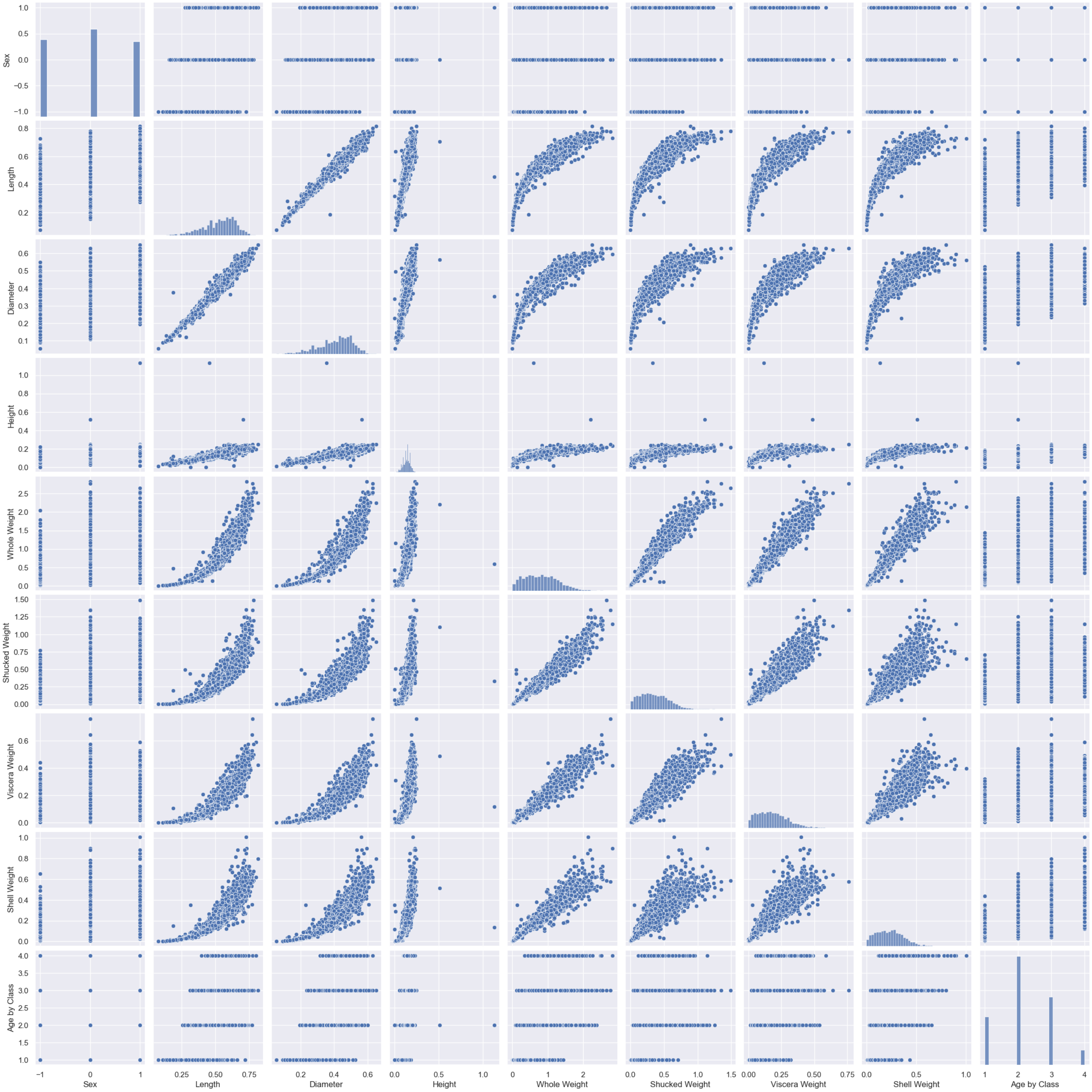

Fig. 1. Pairplot of modified data.

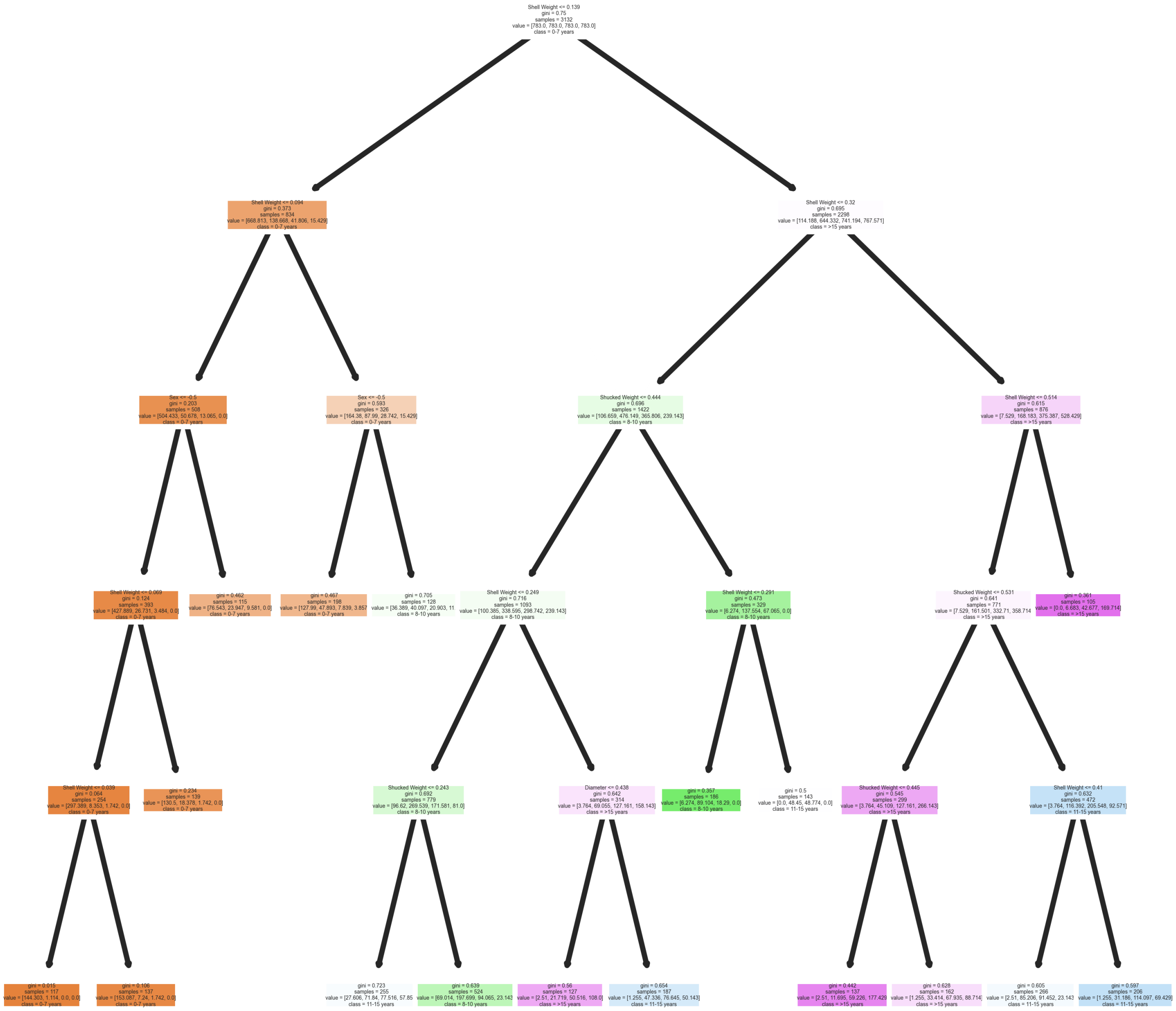

Fig. 2. Original CART tree.

TABLE II
SAMPLES PER AGE CLASS

	Age by Class
2	1891
3	1186
1	839
4	261

The training accuracy of the improved model is 66.96%, with a testing accuracy of 56.08%, which represents a marginal improvement.

C. Random Forest Classification Model

A random forest was created to improve upon the single classification model. The model had the following parameters:

```
{'bootstrap': True, 'ccp_alpha': 0.0,
'class_weight': 'balanced',
'criterion': 'gini',
’max_depth’: 5, ’max_features’: 3,
’max_leaf_nodes’: None, ’max_samples’: None,
’min_impurity_decrease’: 0.0,
’min_samples_leaf’: 100,
’min_samples_split’: 2,
’min_weight_fraction_leaf’: 0.0,
’n_estimators’: 5, ’n_jobs’:
None, ’oob_score’: False,
’random_state’: RandomState(MT19937) at
0x2BCDE2940,
’verbose’: 0, ’warm_start’: False}
```

The model had a training accuracy of 55.62% and a testing accuracy of 55.12%. Performance is comparable to the initial single classification model.

A grid search was performed on the random forest to discover the best set of hyper-parameters. The best estimator had the following attributes:

```
RandomForestClassifier(
class_weight=’balanced’,
n_estimators=44,
random_state=RandomState(MT19937) at
0x2BCDE2440)
```

This model had a training accuracy of 100%, and a testing accuracy of 64.5%. The model outperforms all others on testing accuracy, but with the risk of higher variance.

D. Further Experimentation

Ten experiments were conducted using different seed values to validate the performance of the above models. The hyper-parameters of the models were selected based on the ’best cases’ as discovered during the grid search for each. The results are displayed in Table III and Table IV. The random forest model has the best average test-set performance, with a mean accuracy of 62%. The results of the random forest model also have less variability, with a lower standard deviation.

TABLE III
CART SUMMARY

Count	10
Mean	0.535694
Std	0.026206

TABLE IV
RANDOM FOREST SUMMARY

Count	10
Mean	0.620287
Std	0.018522

IV. CONCLUSION

Classifying the age of Abalone using readily-available physical attributes provides moderate results with high consistency using the random forest methodology. Single CART classification yields inferior performance. Further increases to performance may be possible with the recording of other Abalone features, or with different methodologies like XGBoost or neural networks.

REFERENCES

[1] Warwick J Nash, Tracy L Sellers, Simon R Talbot, Andrew J Cawthorn and Wes B Ford (1994) ”The Population Biology of Abalone (Haliotis species) in Tasmania. I. Blacklip Abalone (H. rubra) from the North Coast and Islands of Bass Strait”, Sea Fisheries Division, Technical Report No. 48 (ISSN 1034-3288)