CS7312: Assignment 1
Gentry Atkinson
Due: 19 February, 2019

1. Introduction: One of the oldest and most prolific forms of classifier within the field of Machine Learning is the Decision Tree. Although many of the newer forms of classifier are able to provide more accurate and flexible results with noisier or less structured data, Decision Trees still maintain a place of prevalence due to the ease with which a human can read and interpret their process of classification. Relative to systems like SVMs, Naive Bayesian classification, and Neural Networks, Decision Trees are extremely simple and intuitive. This makes them the "white box" solution where easily understood results are desired.

Amongst computer scientists in the field of Machine Learning it is generally understood that shorter trees are faster, more efficient, and easier to understand intuitively. This makes algorithms which generate shorter trees without sacrificing accuracy very desirable. This can be done either by generating a complete tree which measures every attribute within an instance and then pruning the branches of that tree which are not contributing substantially, or by carefully weighing each attribute to determine its importance before it is included in the tree.

The purpose of this paper is to compare two common algorithms for generating decision trees: C4.5 and Random Tree. The comparison is being done to determine which tree will generate the shortest tree while maintaining the highest accuracy of classification. More about these algorithms will be discussed later in the paper.

2. Data Sets: The data for these experiments were all taken from the UC Irvine: Center for Machine Learning and Intelligent Systems archives published at http://archive.ics.uci.edu/ml. These data are curated to provide good sets for training and testing Machine Learning Platforms. All data was converted into CSV format before processing in order to improve portability but was otherwise unchanged. 11 sets were chosen from the archive.

2.1 Balance Scale Data Set: Generated in 1976 and posted to the archive by Tim Hume. This data set contains 625 instances representing a balanced scale which has two weights, each some distance from the fulcrum. The attributes of the instance represent the weight on the left and right sides, and the distance each weight is from the fulcrum. Each instance is classified as being left-leaning, right-leaning, or balanced.

2.2 Banknote Authentication Data Set: Generated in 2012 and posted to the archive by Helen Doerken. This set contains 1,372 instances of processed banknotes. Each instance has attributes representing the variance, skewness, curtosis, and entropy of the scanned image of a banknote. The instances are classified as being authentic or inauthentic.

2.3 Car Evaluation Data Set: Generated in 1990 and donated by Marko Bohanec and Blaz Zupan. This set contains 1,728 instances each representing a vehicle. Each instance has attributes representing the

price, maintenance demands, number of doors, number of passengers, lug boot, and safety rating of each vehicle. The instances are classified as unacceptable, acceptable, good, or very good.

2.4 Connect 4 Data Set: Generated in 2003 and donated by John Tromp. This set contains 67, 557 instances representing games of Connect 4. The goal of this game is for a player competing against another to drop tokens into a 6x7 board in such a way that 4 of the player's tokens align. The attributes of each instance describe the state of every position of the 42 possible positions of a board with rows 1-6 and columns a-g. Each instance is classified as a win, a lose, or a draw.

2.5 Contraceptive Method Choice Data Set: Generated in 1987 and donated by Tjen-Sien Lim. This set contain 1,473 instances each describing a married couple who has provided information to a surveyor. The attributes of each instance describe the Wife's age, the Wife's education level, the Husband's education level, the number of children born to the couple, the Wife's religion, the Wife occupational status, the Husband's occupation, their standard of living, and their media exposure. Each instance is classified as "no use", "long-term method", or "short-term method".

2.6 Fertility Diagnosis Data Set: Generated in 2010 and donated by David Gil and Jose Luis Girela. This set contains 100 instances representing donors of semen samples to the World Health Organization. Samples were diagnosed as being fertile or not. The attributes of each instance represent the season the the analysis was performed, the age of the volunteer, whether the volunteer suffered certain childhood diseases, whether the volunteer suffered accident or trauma, whether the volunteer has had surgical intervention, the occurrence of a high fever within a year, frequency of alcohol consumption, smoking habits, and hours spent sitting daily. Each instance is classified as normal or altered.

2.7 Forest Type Mapping Data Set: Generated in 2012 and donated by Brian Johnson. This set contains 523 instances of remote sensing data gathered by the ASTER imagery satellite. The attributes of each instance describe the electromagnetic signature of a single sensor sample. The instances are classified with character representations of various forest types (d, s, o, and h).

2.8 Glass Identification Data Set: Generated by 1987 and donated by Vina Spiehler. This contains 214 instances with each representing a measurement of the chemical signature of a sample of glass. The attributes represent the refractive index, sodium content, magnesium content, aluminum content, silicon content, potassium content, calcium content, barium content, and iron content of each sample. The samples are classified by the application: building float processed, building non-float processed, vehicle float processed, vehicle non-float processed, container, tableware, and headlamp.

2.9 Ionosphere Data Set: Generated in 1989 and donated by Vince Sigillito. This data set has collected samples of observations of the ionosphere by ground based radar systems. Each instance has collected 17 pulse numbers with each number represented by two continuous numerical values. The instances are classified as good or bad as determined by whether the ionosphere will allow easy communication.

2.10 Qualitative_Bankruptcy Data Set: Generated in 2014 and donated by by A. Martin, J. Uthayakumar, and M. Nadarajan. The set contains 250 instances with each instance representing the

economic health of a single firm. Each firm is described as positive, average, or negative in terms of their industrial risk, management risk, financial flexibility, credibility, competitiveness, and operating risk. Each instance is classified as bankrupt or not bankrupt.

2.11 Urban Land Cover Data Set: Generated in 2013 and donated by Brian Johnson. This set contains 168 instances with each instance representing an aerial photograph of an urban area. The intent of the researchers was to identify the variety of ground cover depicted in the photograph. The 148 attributes of each instance describe the photograph in terms of area, brightness, spectral values density, etc. Each instance is classified as trees, grass, soil, concrete, asphalt, buildings, cars, pools, or shadows.

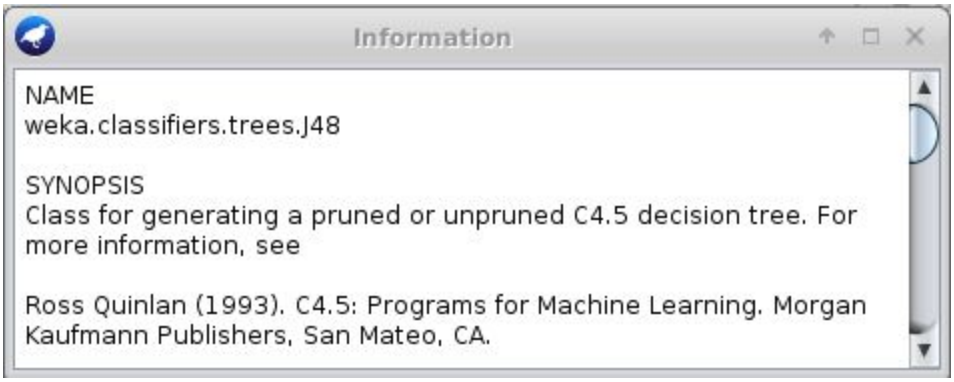
	# Instances	# Attributes	# Classes
Balance Scale	**625**	**5**	**3**
Banknote Authenticity	**1372**	**5**	**2**
Car Evaluation	**1728**	**7**	**4**
Connect 4	**67557**	**43**	**3**
Contraception Choice	**1473**	**10**	**3**
Fertility Diagnosis	**100**	**10**	**2**
Forest Types	**523**	**28**	**4**
Glass Types	**214**	**10**	**7**
Ionosphere Condition	**351**	**35**	**2**
Qualitative Bankruptcy	**250**	**7**	**2**
Urban Land Cover	**168**	**148**	**9**

This chart collects the "size" of all 11 data sets in terms of total entries and dimensionality.

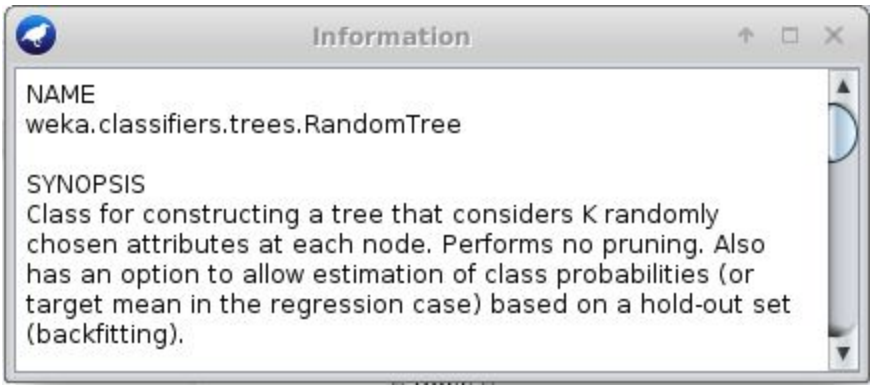
3. Algorithms:

This paper has selected two algorithms for comparison: C4.5 and Random Tree. In both cases the employed implementation is provided by the Weka 3.8.3 machine learning platform, produced and provided by The University of Waikato.

3.1 C4.5: The C4.5 algorithm expanded on the older ID3 algorithm. Decisions made by these trees are determined by the decrease in entropy that results from a split in the data sample. The J48 class provided in the Weka package implements the C4.5 algorithm. All experimental trees were generated with pruning turned on.



3.2 Random Tree: This algorithm considers a random collection of attributes from those which are available in an instance. Each experimental tree was generated using parameters that would consider a number of attributes relative to the log base 2 of the number of parameters.



3.3 k-Fold Cross Validation: This is a method of classifier testing which repeatedly tests the classifier using a varying testing set. Briefly, the full set of instances is divided into k subsets, each containing (1/k) instances. Training is performed using (k-1) of the groups and testing is performed on one group. Training and testing proceed while rotating the one test group until every group has been used as the test set. Furthermore, the groups can be stratified to ensure a representative portion of classes be present in each group.

All experiments in this project have been performed using stratified, 10-fold validation. So each round is performed using 90% training data, and 10% testing data. This process repeats 10 times.

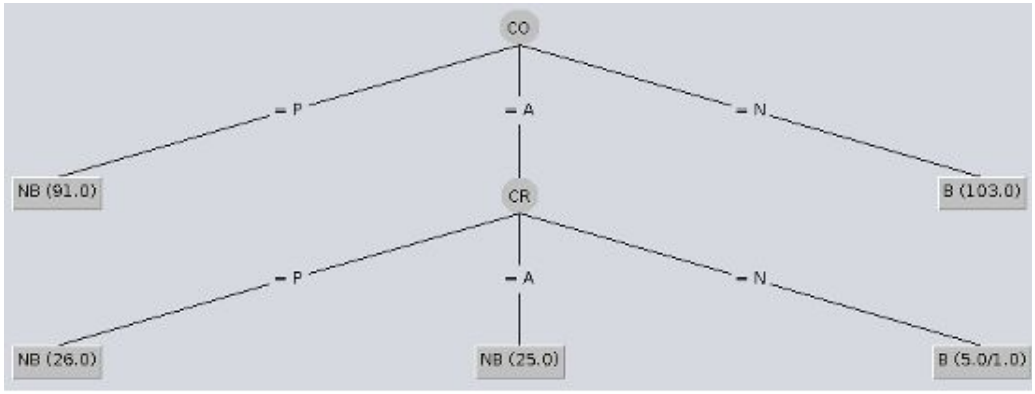
4. Results: This project has demonstrated that pruned C4.5 are in fact shorter that Random Trees. Furthermore, the reduction in tree height occurs without diminishing the accuracy of the classifier. Conversely, the J48 class was generally able to classify the testing data with a greater degree of accuracy that was the Random Tree algorithm.

	J48 Height	J48 Correctly Classified	J48 Precision	Random Height	Random Correctly Classified	Random Precision
Balance Scale	8	76.64%	0.732	12	78.40%	0.815
Banknote Authenticity	8	98.54%	0.985	7	98.76%	0.988
Car Evaluation	6	96.35%	0.965	8	93.34%	0.934
Connect 4	14	80.90%	0.791	18	70.33%	0.696
Contraception Choice	4	48.88%	0.478	5	46.50%	0.462
Fertility Diagnosis	1	85.00%	0.771	12	81.00%	0.841
Forest Types	9	87.57%	0.876	14	84.32%	0.844
Glass Types	10	65.89%	0.658	13	69.63%	0.706
Ionosphere Condition	11	91.45%	0.915	12	87.75%	0.877
Qualitative Bankruptcy	2	98.00%	0.98	4	98.80%	0.988
Urban Ground Cover	7	76.17%	0.8	9	62.50%	0.636
Arithmetic Mean	7.27	82.31%	0.81	10.36	79.21%	0.80
Deviation	3.69	14.41%	0.15	3.98	15.34%	0.15

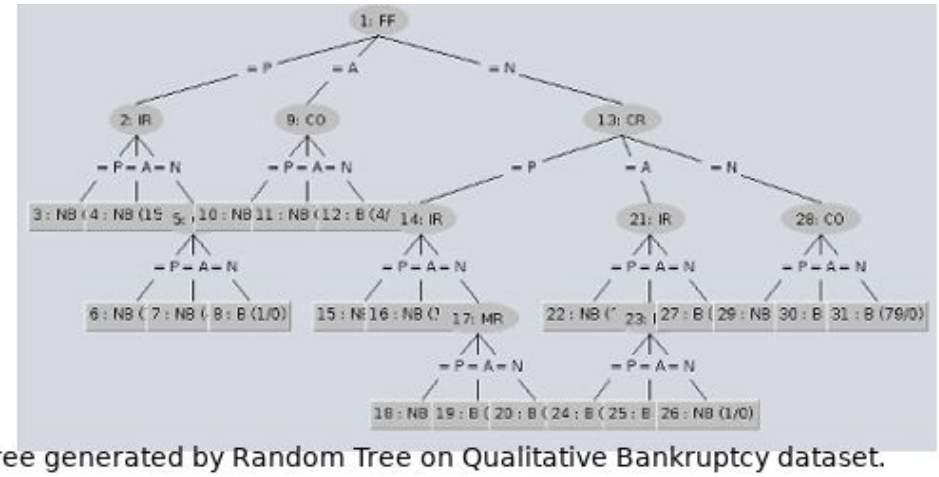
This chart gives the height, accuracy, and precision of J48 and Random Trees across 11 experiments. Green highlights indicate highest accuracy in a row. Yellow highlights indicate the highest precision in a row.

As the above chart demonstrates the J48 trees were more accurate in seven out of the eleven experiments. Furthermore, J48 had a higher precision in six of the eleven test cases. We can also see that the average tree height is about 7.3 tiers for the J48 against an average of 10.4 tiers for the Random Trees. To fully understand the implications of this difference, we need to understand that the size of trees grow exponentially relative to their number of tiers. So a full binary tree with three extra tiers will have eight times as many nodes as a full binary tree with three fewer tiers. These numbers indicate that a Random Tree will be substantially larger than one generated by the C4.5 algorithm.

This does not mean that C4.5 invariably produces a shorter tree. We can see in the second experiment, Banknote Authenticity, that the C4.5 tree is one tier taller. Furthermore the Random Tree is slightly more accurate and precise as well as being shorter in this case. Nonetheless, on average C4.5 trees are shorter.



Tree generated by J48 on Qualitative Bankruptcy dataset.



Tree generated by Random Tree on Qualitative Bankruptcy dataset.

The above images also demonstrate that a well pruned is also more likely to be regular and well balanced. This regular formation can greatly contribute to the ease of human interpretability of a tree. Since this is one of the most desirable features of Decision Tree classifiers, we can see that a well pruned tree like that produced by the C4.5 will be greatly desirable. Any leaf node in the top tree can be

expressed in terms of only two attributes. By contrast some leaf nodes in the lower tree can only be described as the result of four distinct decisions, making the outcomes of that tree more difficult to understand by human entities.

5. Conclusion: This paper has demonstrated that decision trees generated by C4.5 are, on average, 3 tiers shorter than Random Trees on 11 different data sets taken from a variety of fields including image analysis, weather analysis, and financial analysis. These data sets were collected over a range of decades and included data of very different dimensionality. This indicates that it is a durable and portable conclusion that C4.5 trees are shorter than Random Trees.

The shorter tree generated by C4.5 were also found to be more accurate and precise in a majority of test cases. Although we cannot conclude that shorter trees are necessarily more accurate, it has been demonstrated here that good pruning strategies can shorten a tree without degrading its accuracy or its precision.