CE888 Assignment 2: Knowledge Distillation from Random Forests

David Barragn Alcntar, *Data Science and Desicion Making, University of Essex,*

***Abstract*—Data Science plays an important role in dealing with big amounts of information. Different methods have been developed to obtain data from databases. An important field in Machine Learning is data classification in order to make intelligent decisions based on class predictors. This report describes the most employed classifier methods making emphasis on random forests, which will be used to learn from three different data sets. Besides, a procedure known as knowledge distillation will be implemented upon the random forests to measure its impact on classification accuracy. Finally, the model developed in this report will be compared to state-of-the-art methods.**

***Keywords*—Data Science, Machine Learning, Random Forest, Classifiers, Knowledge distillation, Classification methods, Decision tree.**

I. Introduction

Data science is a multi-disciplinary field of computer science combining statistics with the analysis of considerable amounts of data and machine learning techniques. The final outcome of these procedures is to generate knowledge and understanding from data sets [1].

As the technology has advanced in the last decades, the quantity of information carried out by institutions and organizations has grown exponentially, leading to consider conclusions that can be generated from the information in the data collected through time.

One of the main objectives of data science is to improve decision-making capabilities in government, business and medicine, just to name a few, thanks to the extraction of information.

A common use of the techniques of data science that are applied in real-world scenarios is in business marketing, such as online advertising, targeted marketing advertising and selling recommendations. Another recurrent application of data science is customer analysis, where financial industries assign a credit score to a client in order to predict whether a person will be able to pay the credit or not [2]

The present report includes a data science and machine learning technique used for classification and regression. The name of this procedure is ”Knowledge distillation from random forest” that basically uses a random forest method to extract and compress information into a single comprehensive and interpretable decision making tree-like model (decision tree). This approach will be explained in more detail in the Methodology section of the present work.

The method implemented will be compared with other actual methods for similar classification/regression algorithms applied to three different sets of data obtained from [3], an online repository containing a wide variety of data sets suitable for machine learning techniques.

The three data sets are focused on classification approaches rather than regression procedures. The reason for this is because the majority of the data sets found in [3] are adequate for this specific task. The data sets will be fully explained in the Methodology section of this report.

After the implementation of the random forest method to the data sets, the extraction of information is expected from the sets of data. This new and previously hidden at first appearance information will be useful for making decisions and future predictions on unseen instances.

II. Background

Due to the present work deals with a classification problem, following there is a description of some classification techniques that have been used in the past, along with a discussion in previous work on the topic.

Classification algorithms belong to the category of supervised learning. This division refers to the methods used to correlate a number of input parameters into an output, generally a categorical value [4].

In order to learn, a classifier needs to somehow generalize the information extracted from a set of example data, where the output class is assigned to a specific set of input values. This example of data is generally called ”training data” [5]. After the model has been trained, another, different and unseen set of data is applied to the algorithm to compare the predicted values of the classifier with the real values. This second set of data is named ”test data”. This is useful to measure the performance of the classifier, the more similarity between the model results and the real outputs, the more accurate the algorithm is, and the more generalization ability it possesses.

According to [6], the most prevalent machine learning algorithms for classification are linear classifiers, support vector machines, decision trees, boosted trees, random forest, neural networks and nearest neighbours. In all methods listed above, the model learns from the data to classify new observations into two or more classes.

A. Linear Classifiers

A linear classifier is one of the simplest ways to group different classes. The mechanics of a linear classifier is to make a decision based on a linear combination of its features (hence its name) [7].

A common problem of linear classifiers arises when the classes are not linearly separable, meaning that the simplicity of the method leads to inaccurate predictions. Thus, more

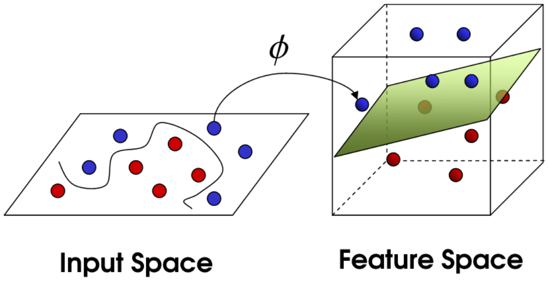

Fig. 1: Support vector machine

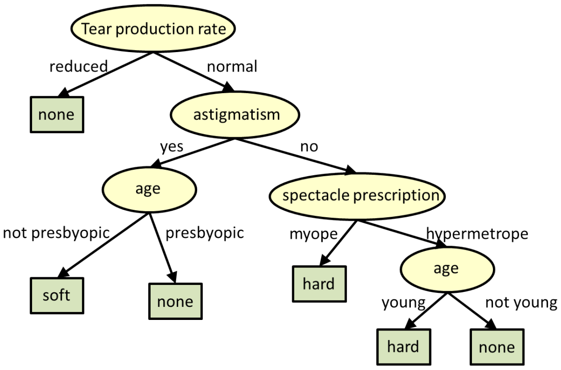

Fig. 2: Decision tree

sophisticated algorithms have been developed in order to deal with this and similar situations.

Some examples of commonly used linear classifiers are naive Bayes classifier, logistic regression and the linear perceptron [8].

B. Support Vector Machine

Support vector machine (SVM) is a classifier and regression procedure that usually separates instances belonging to one of two classes (binary classifier). SVMs are capable of performing linear and no linear classification with the correct use of the kernel, a mathematical tool that allows an SVM to map the data into higher dimensions where the instances will be separable with linear hyper-planes (Fig. 1) [9].

One characteristic of SVMs is that is a deterministic algorithm, meaning that with the same training data and parameters, the classifier will perform in the same manner. Another property of this method is that it only uses the most prominent data, ignoring and saving resources for no relevant information [10].

C. Decision Tree

A decision tree is a method generally used to classify instances using a branching approach to separate the different classes according to their key features. A decision tree classifier is a tool that uses a tree-like model consisting of nodes, branches and leaf nodes. Inside each node, a feature is measured and branched into two or more paths, until there is only one class in the leaf node or a stop criteria has been reached (Fig. 2) [11].

The important step in building a decision tree is the feature selection for each node. A good feature selection will lead to smaller trees, whereas a bad selection of these features can provide no information valuable to classify.

An advantage of a decision tree is that it can be easily understandable, and can be represented graphically as well. Each node represents explicitly a decision-making process [12].

D. Boosted tree

Gradient boosting is a technique used for classification and regression in supervised learning. This method makes use of weak decision tree predictors, combining and converting them into a stronger decision tree through ensembling them [13].

The term ”boosted” refers to the method of which the weak predictors become stronger through several iterations, In each iteration, the last decision tree predictions are used to generate the next classifier [14].

E. Random Forest

A random forest is a similar approach to a boosted tree classifier. Both models share the dependency on weak decision trees to combine them and construct a stronger one. The difference with the random forest is the method to generate the weaker decision trees and the procedure to merge them into a better tree classifier.

Random forest is also an ensemble learning method. It first produces a set of randomly generated decision trees with randomly selected features during the training period, hence the name. Then, for constructing the final decision tree classifier, the method uses the mode of the classes of all individual trees (Fig. 3) [15].

An advantage of this method over single decision trees is that a random forest prevents over-fitting on the training set, achieving more generalization and accuracy on the test set [16].

Another quality of the random forest is that it allows measuring the feature importance of the overall random forest by examining the same feature across all single decision trees and how that selection led to a decrease in the impurity in the forest [17].

Apart from the advantages mentioned above, the random forest has the drawback of the need for a big number of trees, causing slowness in the model and converting it an ineffective real-time classifier. Generally, because of the random generation of trees in the training session, the algorithm is faster to train than to actually predict on the test set. Nevertheless, for real-world applications, the random forest is fast enough to handle the majority of classification tasks [17].

F. Neural Network

A neural network, also know as an artificial neural network (ANN) to differentiate it from the human nerve system, is a

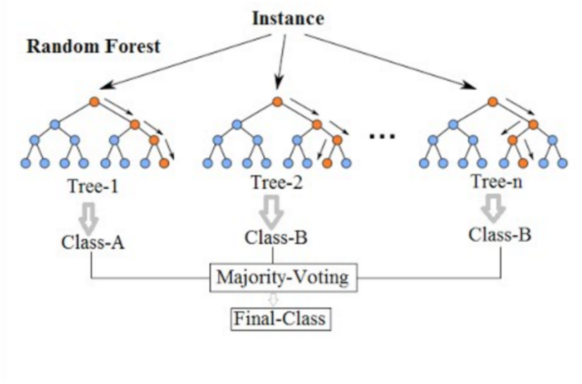

Fig. 3: Random forest

model based on the interconnection and flow of information in biological neurons.

ANNs learn from examples and adjust their parameters to generate the same or a near output as the training data. They are used for classification and regression as well. An ANN is composed of interconnected nodes called ”perceptrons” that perform mathematical operations with the outputs they receive and produce an output passed to subsequent perceptrons (Fig. 4) [18].

Generally, the more perceptrons a neural network has, the more power of generalization it achieves. Furthermore, there are several classifications of neural networks depending on their interconnection. ANNs are widely used in image classification, speech recognition, medical diagnosis and game playing [19].

All methods described above are commonly used for classification purposes. For the intention of this report, the random forest algorithm will be selected to classify three different sets of data obtained from [3] and the results will be compared with previous work on the same data sets. The following section describes in depth the selected method as well as the information in the data sets.

III. Methodology

The goal of the present work is a supervised learning classification method of three different sets of data. The outcome will be a random forest classifier that will allow to a pure decision tree classifier to extract or distil knowledge ”hidden” in the data sets.

Firstly, the three data sets will be described, this information was obtained from [3], a machine learning repository with more than 400 different data sets for machine learning, data science and data mining purposes. The first data set selected corresponds to the ”Abalone Data Set” to predict the age of Abalone shellfishes. The second data set is the ”Human Activity Recognition Using Smart-phones Data Set” which classifies six different human behaviours. Finally, the third data set is called ”Adult Data Set” and is used to predict the income of a certain person.

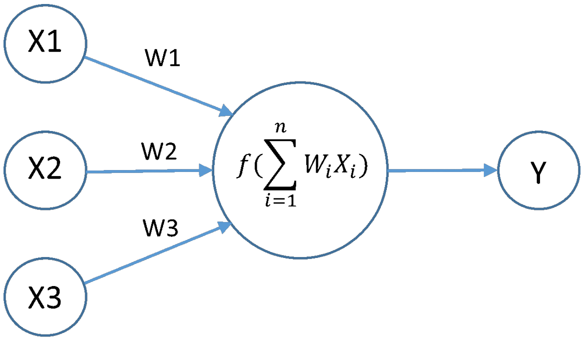

Fig. 4: Perceptron of a neural network

A. Abalone Data Set

The Abalone data set is used to predict the age of abalone shellfishes depending on physical measurements. The age of a single abalone is measured by cutting its shell and counting the number of rings with the help of a microscope. The data comes from a population study of abalone described in [20].

The data set contains information of 4,177 instances, each one with 8 different features that can be categorical, composed of natural numbers or real numbers. The following list shows the attributes of this data set.

- Sex (M / F / Infant)
- Length (mm)
- Diameter (mm)
- Height (mm)
- Whole weight (grams)
- Shucked weight (grams)
- Viscera weight (grams)
- Shell weight (grams)

- Output: Rings (+1.5 gives the age in years)

B. Human Activity Recognition Using Smart-phones Data Set

Human Activity Recognition data set is used to classify a person’s behaviour using an accelerometer on a smart-phone located on the waist. The data was generated with a group of 30 volunteers from 19 to 48 years old. Each person performed six actions (walking, walking upstairs, walking downstairs, sitting, standing and laying) [21]

The data set contains information of 10,299 instances, each one with 561 attributes. The following list shows the main features of this data set.

- Triaxial acceleration from an accelerometer
- Triaxial angular velocity from a gyroscope
- 561-feature vector with time and frequency domain variables.
- Identifier of the subject

- Output: Activity label

C. Adult Data Set

The Adult data set is used to predict whether the income of a person exceeds \$50,000 per year based on census data. The data was extracted by Barry Becker from the 1994 census database [22].

The data set contains information of 48,842 instances, each one with 14 different features that can be categorical or composed of natural numbers. The following list shows the attributes of this data set.

- Age (continuous)
- Work-class (categorical)
- Fnlwgt: (continuous)
- Education (categorical)
- Education-num (continuous)
- Marital-status (categorical)
- Occupation (categorical)
- Relationship (categorical)
- Race (categorical)
- Sex (categorical)
- Capital-gain (continuous)
- Capital-loss (continuous)
- Hours-per-week (continuous)
- Native-country (categorical)

- Output: $>$\$50K/yr, $\leq$\$50K/yr,

These data sets were selected because of their instances size, the association task (classification) and the different types of data (continuous, categorical, time series). Furthermore, the classification tasks are binomial (in the case of the Adult Data Set) and multi-class (in the case of the Abalone Data Set and the Human Activity Recognition Using Smart-phones Data Set). These differences between data sets will help to identify if the knowledge distillation approach behaves similarly with distinct sets of data.

D. Method selected

Regarding the method for analyzing the data, the model that will be used to classify and predict the outcomes of the three different data sets will be a random forest.

Random forest was selected because of its robustness due to the number of trees involved in the training and predictions. Furthermore, it does not suffer from overfitting for the reason that it averages all the predictions of simple random decision trees cancelling any favouritism towards certain results.

The output of the random forest will not be the final predicted class, instead, the model will provide the probabilities for every class (being the maximum probability the final result). This set of probabilities will be used to create a similar dataset from the original, with the difference that the output, instead of being just one category, will be the probabilities of being each one of the different classes.

The second dataset (with the probabilities as the output) will be trained on a decision tree. This tree will hence predict not a single category, but the probabilities of every category, trying to mimic the same behaviour as the random forest classifier. Once obtaining the different probabilities from the test set, the decision tree will select the highest probability to provide the final classification.

This overall process of using a simpler model to mimic a more complex behaviour through class probabilities is what is known as Knowlege Distillation. The advantage of distilling knowledge is that an uncomplicated algorithm can achieve a near performance as a more sophisticated algorithm. Furthermore, the principle of the distillation resides on the information of the classes different from the maximum probability carrier, providing more ”knowledge” about the relations between classes.

In order to compare the performance of the decision tree distillator, a raw decision tree with the same characteristics in size and shape will be implemented on the original data set (without the probabilities). It is expected that the knowledge distillation provides better performance on the decision tree.

E. Tools for the analysis of data

For the development of the random forest and decision tree models, the scikit-learn python toolkit will be employed [23]. This toolkit has a built-in function called ”RandomForestClassifier”, ”DecisionTreeRegressor” and ”DecisionTreeClassifier” that performs the methods described previously. The classifiers are provided with various parameters that can be modified to produce classifier models with different characteristics, for example, the number of estimators (in the case of the random forest), the maximum depth and the criteria for the split of the trees, among other parameters. The DecisionTreeClassifier will be implemented as the raw decision tree, whereas the DecisionTreeRegresor will be used as the decision tree distillator due to the DecisionTreeClassifier cannot provide multi-output for the different probabilities to match. However, the same operating principles apply for both decision trees.

IV. Experiments

In order to perform the experiments, a series of preprocessing tasks were implemented on the data sets. Due to the methods for the experimentation deal with numbers and not categories, all non-numerical values were encoded into integer numbers. For example, in the Abalone Data Set, the feature ”Sex” contains values for masculine (M), feminine (F) and indifferent (I). Each category was encoded as M-2, F-0 and I-1 (following an alphabetic order).

For the Abalone Data Set and the Adult Data Set, the data was split into training data and testing data with a ratio 7:3 respectively. The Human Activity Recognition Using Smartphones Data Set was not split because it already provides both sets, training and testing.

After the pre-processing of the data, all the following experiments were performed equally on the three data sets. Firstly, a random forest classifier was trained using 250 estimators. After the training phase, the class probabilities for each instance in the training set was generated. Figure 5 shows the class probability distribution for the first instance of the Abalone training set.

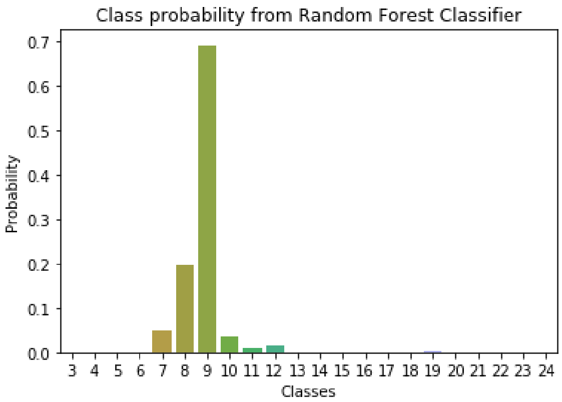

Fig. 5: Class probability distribution from random forest classifier on the first instance of the Abalone training set

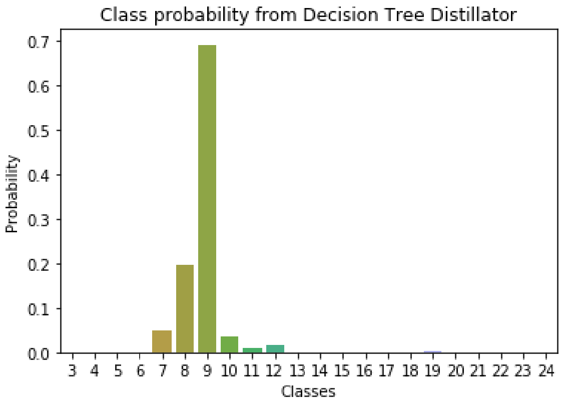

Fig. 6: Class probability distribution from decision tree distillator on the first instance of the Abalone training set

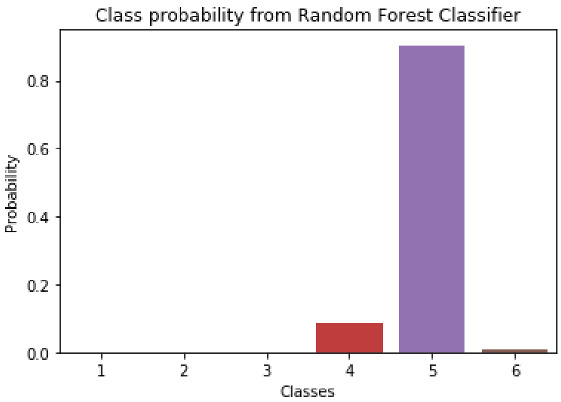

Fig. 7: Class probability distribution from random forest classifier on the first instance of the Human Activity training set

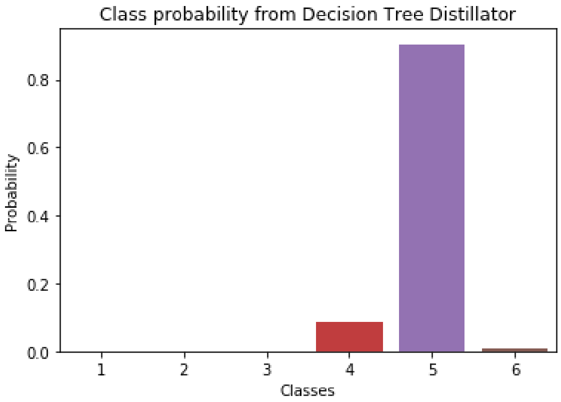

Fig. 8: Class probability distribution from decision tree distillator on the first instance of the Human Activity training set

The next step is to produce a decision tree distillator that behaves as the random forest classifier, producing similar probability distributions. Figure 6 shows the probabilities for the same instance of the Abalone training set as Figure 5 but with the predictions of the decision tree distillator.

Similarly, Figure 7 and Figure 8 represents the probability distributions of the Human Activity Recognition Using Smart-phones Data Set using the random forest and decision tree distillator respectively. The same approach is achieved in Figure 9 and Figure 10 on the Adult Data Set.

An advantage of both, random forest and decision trees, is that it is possible to measure the feature importance for each data set. In tree-like algorithms, the most important feature at every point of the tree is selected to perform further subdivisions. Hence, each feature receives an importance score proportional to the contribution to the final classification. The most important features play a prominent role in the classification, whereas the less important features may be not even related to the final outcome.

Table I list the five most important features from the Abalone Data Set from the most important to the less important, depending on the model implemented.

Similarly, Table II shows the five most important features from the Human Activity Recognition Using Smart-phones Data Set (with numbers as there are 561 different features). And Table III lists the most predominant features of the Adult Data Set by importance order.

In order to measure the performance of the three different

Imp	Random Forest	Decision Tree Distillator	Raw Decision Tree
1	Shell weight	Shell weight	Shell weight
2	Shucked weight	Shucked weight	Shucked weight
3	Whole weight	Viscera weight	Viscera weight
4	Viscera weight	Diameter	Whole weight
5	Length	Whole weight	Diameter

TABLE I: Feature importance for the Abalone data set depending on the model implemented

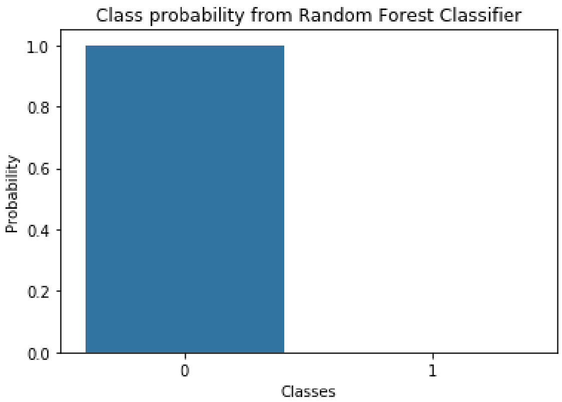

Fig. 9: Class probability distribution from random forest classifier on the first instance of the Adult training set

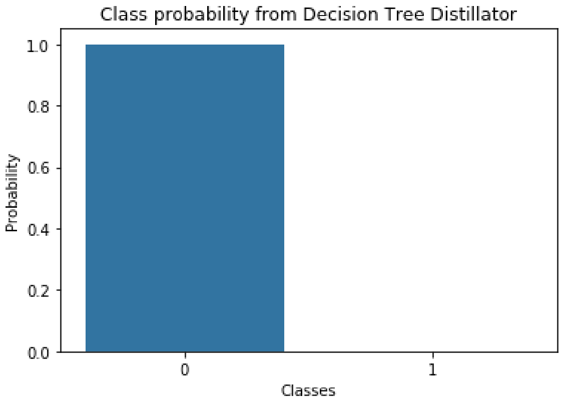

Fig. 10: Class probability distribution from decision tree distillator on the first instance of the Adult training set

methods, a series of one hundred iterations were performed on each data set to average the accuracy, which was calculated using the accuracy score provided by the ”metrics” library from the sklearn toolkit.

Figure 11 plots the average accuracy of the three different models through all the iterations. On each iteration, the average accuracy is calculated adding all previous accuracies and dividing them over the number of iterations (aka arithmetic mean).

Figure 12 and Figure 13 show the plots of the average accuracy through one hundred iterations for the Human Action Data Set and the Adult Data Set respectively.

Imp	Random Forest	Decision Tree Distillator	Raw Decision Tree
1	52	52	52
2	56	271	389
3	558	559	559
4	41	508	508
5	49	74	74

TABLE II: Feature importance for the Human Action data set depending on the model implemented

Imp	Random Forest	Decision Tree Distillator	Raw Decision Tree
1	Fnlwgt	Relationship	Fnlwgt
2	Age	Education-num	Relationship
3	Capital-gain	Capital-gain	Age
4	Relationship	Fnlwgt	Education-num
5	Education-num	Age	Capital-gain

TABLE III: Feature importance for the Adult data set depending on the model implemented

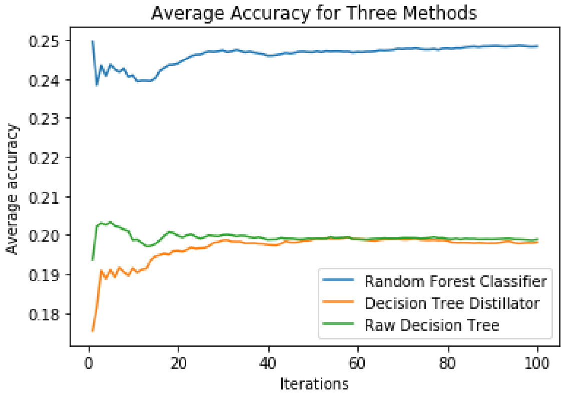

Fig. 11: Average accuracy for the Abalone data set using three different models

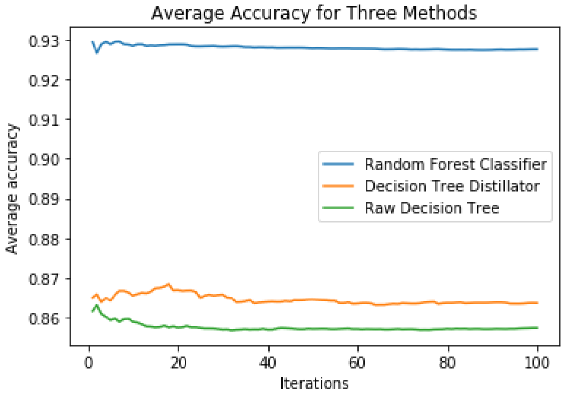

Fig. 12: Average accuracy for the Human Action data set using three different models

Finally, Table IV compresses the final accuracies from the three data sets using the three different models.

V. Discussion

Starting with the Abalone dataset, Figure 5 replicates the manner a random forest classifier decides which class the instance belongs to. On this specific case, class 9 contains the major probability (approx. 70%), whereas the second most probable outcome has a probability value of around 20%.

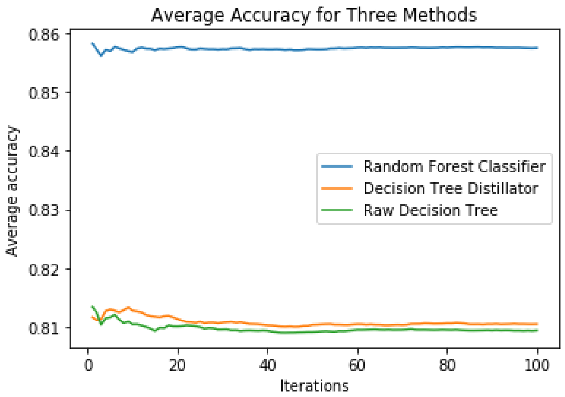

Fig. 13: Average accuracy for the Adult data set using three different models

Model	Abalone	Human Action	Adult
Random Forest Classifier	0.2484	0.9276	0.8574
Decision Tree Distillator	0.1981	0.8635	0.8104
Raw Decision Tree	0.1989	0.8573	0.8094

TABLE IV: Average accuracy for three data sets with the implementation of three different models

Therefore, it is expected that the result will be 9 rings, as the number of rings is the final outcome for each instance.

The distillation process can be seen clearly in Figure 6, where a much simpler decision tree was trained to predict the same probabilities as the random forest classifier. It can be seen that the major probability is still in class 9 with approximately the same value of 70%, and the second most probable class remains the same with about 20%. Visually, there are imperceptible differences as the shape of the bars resembles almost the same.

For the Human Action dataset, the major class probability represented in Figure 7 is category 5 with a value above the 80%, and the second most probable class with a value below 20%. Figure 8 reassures the similitudes between the decision tree distillator and the random forest classifier. The similar case occurs with Figure 9 and Figure 10 for the Adult data set, where there is a binary output (only two possibilities). The figures show a hundred per cent in probability for the specified instance.

The considerable difference between the most probable value and the second most probable value relies on that the distillation process was performed on training data, thus, the algorithm is almost certain that the output will be the same as the one seen in previous training examples. Testing data cannot be used due to it will be needed to measure the accuracy of the three methods, the random forest classifier, the decision tree distillator, and an additional raw decision tree with the same characteristics as the previous tree but without the distillation process.

The next evaluation was the feature importance for each one of the data sets. In the case of the Abalone data set, the three different models have the same most important feature, which is ”Shell weight”. Furthermore, the importance of this feature has a value of around 20%. This means that the shell weight is a strong indicator of the number of rings in an Abalone shell. The three models also share the second most prominent feature, which is ”Shucked weight” with a value of importance around 15%, contributing significantly to the final prediction of the number of rings. From the third position onwards, the features vary from the position but they are almost the same, ”Whole weight”, ”Viscera weight” ”Diameter” and ”Length”. All remaining features have an importance value of 10% or less, meaning that they do not contribute in the same way that the previously mentioned features do.

The feature importance evaluation for the Human Action data set provides no consistent results because this data set contains a total of 561 different features labelled numerically. For the random forest classifier, the most important feature was the No. 52 with an importance value of 3.6%, this low value means that the features are almost evenly distributed and there is no considerable most important feature. On the other hand, the other two tree classifiers have as the most important feature the No. 52 too, but with an elevated importance value of 25% and 24% respectively. This inconsistency signifies that the random forest is trying to equalize the importance of every single feature on this specific data set, whereas the simple decision trees tried to choose a prominent feature from the beginning in order to start dividing its branches. Nevertheless, they only coincide in the first value, the decision tree methods also have in their first five features the No. 559, No. 508, and No. 74, but none of this is present in the random forest classifier.

From the evidence discussed previously, it can be concluded that although the decision tree distillator tries to mimic the behaviour of the random forest through knowledge distillation, It resembles more like a raw decision tree with slightly improvements

For the Adult data set, the three different methods share the same five most important features, but different from expected, they order them in a different manner independently with the highest value around 20% and the consecutive ones falling in an exponential-like way. One possible reason for this is because the dataset has a binary output, in that case, is not that important which one of the five most important features goes first, due to by the fifth feature they will have certainly almost the same answer.

Regarding the final accuracy, Figure 11, Figure 12 and Figure 13 show the plots over a hundred iterations. On all of them, it can clearly be perceived that the random forest classifier outperforms the accuracy of the other two decision trees, with 4%-7% more accuracy than the remaining methods.

Comparing the two decision tree algorithms, the distillator and the raw tree, they seem to behave equally for the Abalone data set (difference in accuracy less than 0.1%) and they have a difference in the accuracy of 0.62% and 0.1% for the Human Action data set and the Adult dataset respectively.

This difference is relatively small but it will perdure in the long run (more iterations) proving that the knowledge distillation algorithm provides slightly better performance when it is applied with a decision tree from the probability outputs of a random forest classifier.

Regarding the overall performance on the three datasets, it can be perceived that on the Abalone data set the algorithms performed poorly compared with the other two data sets. The reason for that is because on the first data set one output has to be selected among almost thirty classes, where there is a correlation between them. In other words, it could not be completely wrong to predict 26 rings when in reality the abalone shellfish has actually 27 rings. Contrasting with the Human Action data set, where the difference of one in the output represents a completely different behaviour. And on the Adult dataset where there are only two possible outputs.

VI. Conclusion

It can be concluded that the knowledge distillation process has a positive effect on the performance of a decision tree classifier when extracting the aforementioned knowledge in the form of probability distributions.

The accuracy of the decision tree distillator outperforms the raw decision tree in about 0.1%-0.7%. But it is still similar to a decision tree when is compared with the original knowledge provider, the random forest classifier.

The random forest classifier by its own performs very well, for that reason, it will be difficult to find a simpler method that has better accuracy. Nevertheless, with this approach, it can be concluded that a simpler method (e.g. decision tree) can be enhanced to improve its performance with the help of the inherent complexity of a random forest by only mimic the class probability distribution behaviour.

Future work will be carried to compare the performance with different data sets and different algorithm structures (e.g. a complex neural network distilling knowledge to a simpler feed forward neural network)

References

[1] V. Dhar, ”Data Science and Prediction”, *Communications of the ACM*, December 2013, Vol. 56, No. 12, Pages 64-73

[2] F. Provost and T. Fawcett, ”Data Science and its Relationship to Big Data and Data-Driven Decision Making”, *Big Data*, February 2013, Vol. 1 No. 1

[3] D. Dua and E. Karra, ”UCI Machine Learning Repository”, *University of California, Irvine, School of Information and Computer Sciences*, 2017, url: http://archive.ics.uci.edu/ml

[4] S.J. Russell and P. Norvig, ”Artificial Intelligence: A Modern Approach”, *Prentice Hall*, Third edition, 2009

[5] M. Mohri, A. Rostamizadeh, and A. Talwalkar, ”Foundations of Machine Learning”, *MIT Press*, Second edition, 2018

[6] M. Sidana, ”Types of classification algorithms in Machine Learning”, *Sifium Technologies*, February 2017

[7] G. Yuan, C. Ho, and C. Lin, ”Recent Advances of Large-Scale Linear Classification”, *Proc. IEEE*, 2012, Vol. 100, No. 9

[8] T. Mitchell, ”Generative and Discriminative Classifiers: Naive Bayes and Logistic Regression”, *Draft version*, 2005

[9] C. Cortes, V.N. Vapnik, ”Support-vector networks”, *Machine Learning*, 1995, Vol. 20, No. 3, Pages 273-297

[10] Scikit-learn, ”Support Vector Machines”, *Journal of Machine Learning Research*, 2011, Vol. 12, Pages 2825-2830

[11] L. Rokach, and M.O. Lior, ”Data mining with decision trees: theory and applications”, *World Scientific Pub Co Inc.*, 2008

[12] J.R. Quinlan, ”Induction of decision trees”, *Machine Learning*, 1986, Vol. 1, Pages 81-106

[13] J.H. Friedman, ”Greedy Function Approximation: A Gradient Boosting Machine”, *IMS 1999 Reitz Lecture*, February 1999

[14] (Electronic Version): StatSoft, Inc. (2013). ”Electronic Statistics Textbook”, Tulsa, OK: StatSoft. WEB: http://www.statsoft.com/textbook/

[15] T.K. Ho, ”Random Decision Forests”, *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, Montreal, August 1995, Pages 278-282

[16] T. Hastie, R. Tibshirani, and J. Friedman, ”The Elements of Statistical Learning”, *Springer*, Second edition 2008

[17] D. Nikla, ”The Random Forest Algorithm”, *Towards Data Science*, February 2018

[18] M. van Gerven, S. Bohte, ”Artificial Neural Networks as Models of Neural Information Processing”, *Frontiers in Computational Neuroscience*, January 2018

[19] D. Silver, et al. ”Mastering the game of Go with deep neural networks and tree search”, *Macmillan Publishers Limited*, 2016

[20] W.J. Nash, T.L. Sellers, S.R. Talbot, A.J. Cawthorn and W.B. Ford, ”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*, 1994, Technical report No. 48

[21] J.L. Reyes, D. Anguita, A. Ghio, L. Oneto and X. Parra, ”Smartlab - Non-Linear Complex Systems Laboratory”, *Universit degli Studi di Genova*, Genoa, Italy

[22] R. Kohavi and B. Becker , ”Data Mining and Visualization”, *Silicon Graphics*

[23] Pedregosa et al., ”Scikit-learn, Machine Learning in Python”, *JMLR*, 2011, Vol. 12, Pages 2825-2830