SCHOOL OF ELECTRONICS AND COMPUTER SCIENCE, UNIVERSITY OF ESSEX

Knowledge Distillation of Random Forest

J. Cebrián

Abstract— High-performance machine learning algorithms can be expensive to store, hard to evaluate, interpret by humans and integrate into systems [1]. The use of simpler models with the same or similar performance as the cumbersome models instead of the actual complex model is the ideal goal for many machine learning applications. The approach of Random Forests (RFs) of using several randomized Decision Trees (DT) that are trained on different samples and random subset of features of the data [2], but that altogether form a reliable model has shown a surprisingly well performance in many other works. That being said, Random Forests are known to be more robust and a stronger technique than a single DT. Nevertheless, on the theoretical side, RFs are hard to interpret and are less conclusive. The latter can be a problem for applications for the health sector were the interpretation of the results and knowing why a choice is taken is of paramount importance for diagnosis and risk prediction. To address this, "Knowledge Distillation" procedures are used to extract knowledge hidden in larger models and apply it to simpler algorithms. The present work assesses the performance of a RF method for three different classification tasks. Then, a form of knowledge distillation is applied where a simpler decision three uses the probabilities yielded by the RF to learn and classify the same data without the problem of computational, time and interpretability cost. At the same time, Random Forest's, the student DT's and a simple DT without distillation procedure's performances on the same test data will be compared as well as with other machine learning methods. It's important to remark that this paper assesses the performance of RF using its probabilities estimations; the latter is important in many areas of medicine such as, surgery, oncology, internal medicine, pathology, pediatrics and human genetics. Many researchers in machine learning and clinicians are collaborating to improve healthcare services and although robust predictions are critical in healthcare research since they are directly related to human lives, the interpretability of the results are also fundamental [3]. This document addresses the performance of the ML techniques taking in consideration a healthcare perspective and uses medical datasets to train and test the models.

Index Terms— Decision Tree, Knowledge Distillation, Machine Learning, Random Forest

1 INTRODUCTION

The size of modern data sets has created a necessity of learning algorithms that can perform with statistical efficiency in spite of the volume of information. Random forests (RF) are a type a supervised learning procedure that are part of the most successful methods currently available. RF essence is the "divide and conquer" principle [4]. This method is an example of an ensemble learner, meaning that it joins the results of an ensemble of simple estimators that prove that the sum of the parts can be greater than the part itself [5]. Random Forests ensemble is built on decision trees (DTs).

Tree based models split data according to certain range values in the features. When splitting, subsets of the dataset are created. The final subsets are called terminal nodes and it contains the prediction of the model using the average outcome of the training data in the node. The algorithm tries to create subsets trying different groupings until it determines the best cutoff per feature and then selects the feature for splitting that would result in the best partition according to its variance and its impurity level. The algorithm loops in this action until a stop criterion is reached [6].

The interpretation of DTs is simple and create good explanations usually called "Human-Friendly Explanations" since they are contrastive and comparable. Nevertheless, non-desirable behaviour of DTs might take place: minimal changes in the input feature can have an important impact on the prediction and can be unstable, meaning that slight changes in the train dataset can create a completely different DT which make them at times unreliable. Also, it is important to note that the interpretability of DT depends on the depth of the tree: The deeper the tree, the more difficult to understand its decision rules and the more over-fitting it might have; that is that instead of providing general solutions, they might be specific to the train dataset or a result of noise properties of the particular data. In other words, individual trees are sensitive to the specific data used to train the tree, and the interpretation of it might be inaccurate if small changes in the training data take place. In this context, the ensemble of various different decision trees that focus on different sample subsets and random subset features of the same dataset improve the performance of the results; this is the function of the Random Forest.

RF add randomness to the model while growing the DTs since they search for the best feature among a random subset of features that result in an increase of diversity. Also, RF prevent overfitting most of the time as it combines the smaller decision trees. In simpler words: Random Forests build multiple decision trees and combines them together to create a more reliable, accurate and stable prediction [7]. After fitting a random forest to training data, it is common to obtain the conditional class probabilities for a test point. This technique has been found to be successful in diverse areas such as medicine, ecology, sports forecasting and many others. [8]

However, Random Forest method has its limitations, for example, RF is a predictive model and not a descriptive tool which can be a burden in certain cases where knowing the reasons behind a prediction is fundamental (E.g. medicine) Also, a large number of trees can make the algorithm slow and ineffective, since in general these algorithms are fast to train but slow to create predictions; Moreover, the greater number of trees to be combined the more

accurate the prediction, meaning that as the user pretends to have a more accurate predictor the model results slower and has a higher computational and storage cost.

One of the objectives of this work is to assess the performance of a Random Forest over a dataset and obtain its classification probabilities. Also, this work aims to achieve a method of knowledge distillation from a random forest. The probabilities obtained from the RF are used as a new dataset to be used by simpler method (i.e. Decision Tree) that will learn from this yielded probability dataset and overcome the instability of the simple DT and also overcome the problem of computational cost and slow performance of having a deep learning method such as RF.

The latter can be achieved by developing "student models" which mimic the predictions of the original model or "teacher". Decision trees can automatically fit high-dimensional functions, reduce computational cost of predictions, facilitate data collections and the final nodes can provide an explanation for the predictions. [9]

The scope of this work is to efficiently compress the knowledge of the random forest containing an ensemble of trees to a single, interpretable tree, by using the method of "knowledge distillation". The relevance of the implementation of this method specifically in this work relies on the possible use of machine learning algorithms to interpret, prevent and predict medical conditions in real-life situations. Clinicians and machine learning researchers often collaborate to develop models that can be reliable to make informed clinical decisions and in most of the cases, deep learning models are less interpretable. The latter can be assessed and solved by applying the methods used in this work.

Through the document a list and brief summaries of previous works in this field will be addressed, as well as the methodology and experiments done (for this and other previous works), to achieve the expected results. Also, a comparison of the Decision Tree without the distilling method, the rainforest itself, and the "student" DT among them and with other machine learning methods will be discussed. As pointed out before, model interpretability for medical predictions is necessary, that being said, feature importance for the final Decision Tree mimic model will be shown. All experiments will be performed using medical data to get the closest scenario to real-life medical situations.

2 Background

In this section, an overview of related works using Random Forest will be provided, and then recent advances in the mimic learning and knowledge distillation approaches will be discussed.

In [10] the use of Random Forests is proposed as a top-notch machine learning method to predict the risk of a medical intervention such as complete knee joint replacement; the article emphasizes in the description and interpretability of machine learning algorithms in the health sector. Similarly, [11] implements RFs to improve the accuracy of predicting Heart disease by using the Cleveland heart disease dataset; the RF accuracy was of 85.81% making it capable of saving many lives by predicting heart failure. In the same way, [12] uses RF as a classifier for microarray data and was shown to be competitive with methods that require fine-tuning or pre-selection of parameters. In the mentioned work the method returns small sets of genes that are not redundant and retains predictive performance. Likewise, a Random Forest implementation resulted in higher accuracy compared to Support Vector Machine (the reference method) to make a computer-aided diagnosis of Alzheimer's disease as seen in [13]. Additionally, after a set of experiments and making a comparison of Random Forests with other methods, [14] finds RF to have the highest accuracy among the classifiers and states that it is a reliable method for healthcare and disease prediction.

As it can be observed from the mentioned literature, and many other works, Machine learning researchers have recognized throughout many experiments and datasets that Random Forests are a robust and high-performance machine learning method. Nevertheless, as it was mentioned before, the computational cost of improving the robustness, reliability and stability of machine learning methods as done in RFs is high. The latter pushes researchers to find new algorithms in which the same essential characteristics are meet but with simpler and less cumbersome methods.

To solve the issue, some works have analyzed the aforementioned method of distilling knowledge from complex algorithms; for example: in [15] is shown that it is possible to compress the knowledge of a neural network by making an ensemble into a single model easier to implement obtaining remarkably well results on MNIST and also proposes to use the method to improve the acoustic model of an Android system. In the same way, [9] considers the use of regression trees as the mimic algorithm for model distillation and states that if a "distillation tree replaces the learned model when making predictions" the tree might be an explanation for how a prediction is made. The latter can be used as a way of making complex model interpretable for health predictions since decision trees have "an intelligible graphical representation that can fit complex high-dimensional functions." [9]. Also, [3] uses Gradient Boosting Trees that mimic the soft targets predicted by the Neural-Network based methods and concludes that the mimic methods could have similar or better response than the teacher method. [3] mentions a "very promising direction for future machine learning research in healthcare domain".

3 Methodology

The analysis that will be performed aims to extract knowledge from a Random Forest using probabilities of the class predicted to train a simpler decision tree that learns from these "soft targets" yielded by the RF.

After successfully training the tree, the performance of the random forest, the decision tree -with and without learning from the RF-, Neural-Network based algorithm and SVM will be compared among them to evaluate their performance. The experiment will make use scikit learn

software in python to train the classifiers and obtain "soft-targets" and Seaborn along with Pandas and NumPy to analyze and visualize the datasets.

3.1 Pseudocode

```
"Knowdledge Distillation" Program {

Input dataset as dataframe

Divide data (Train, validation , test) {
        [Xtrain ytrain] train_test_split
        [Xval yval]
        [Xtest ytest]
}
Train Random Forest on Train subset{
        state the clfRF=RandomForestClassifier
        fit classifier to data clfRF.fit(Xtrain,y train)
}
Use Random Forest with Test & Validation subset {
        Get Classes {clfRF.predict(XTEST}
}
Get Probabilities {
        clfRF.predict_proba(XTEST)
}
Compare accuracy
Visualize random forest
Get feature importance
Use Random Forest in all of the original dataset to get probabilities

Make probability bins {
        Numpy hist (choose number of bins)
}
Create new multivariate dataset {
        Multivariatedataset []

}
Divide multivariatedataset (Train, validation , test){
        [Xtrainmv ytrainmv] train_test_split
        [Xvalmv yvalmv]
        [Xtestmv ytestmv]
 }
Train Decision tree on multivariatedataset {
        ClfDT=Decisiontreeclassifier
        ClfDT.fit(Xtrainmv,ytrainmv)
}
Use Distilled DT with Test & Validation subset {
        Get Classes {clfDT.predict(Xtestmv)
}
        Compare accuracy with ytestmv

        Visualize Decision Tree

Test Decision Tree on original dataset

Binarize the DT classification {
         reconstruct original labels of datasets
}
Compare accuracy of random forest
Train simple decision tree on original dataset
Predict with simple decision tree
Compare accuracies
Compute confusion matrix
Train and Test SVM
Train and Test NN
Compare Machine Learning Methods
}
```

3.2 Dataset description

Three datasets will be used to perform the experiment: The first is originally a multivariate dataset that contains EEG Eye state data [16] that classifies as 1 or 0 depending the EEG signals. '1' indicates the eye-closed and '0' the eye-open state. All data is from one continuous EEG measurement with emotive EEG Neuroheadset and the eye state was detected via camera during the EEG and added manually. The data set consists of 14 EEG values and a value indicating the eye state with 14980 samples.

The second dataset is the Epileptic Seizure Recognition Data Set [17]. (Also a EEG recording) The data-sed is a re-shaped and re-structured version of another dataset; now it consists of 11500 samples with 178 data points (columns) with the last one representing the label as {1,2,3,4,5} (Only "1" represents epileptic seizure) For terms of simplicity, the target will be changed manually to a binary labeling: "1" a present epileptic seizure and "0" not epileptic seizure present.

The last dataset is Immunotherapy Dataset [18]. It contains information about wart treatment results of 90 patients using immunotherapy. The dataset has 8 number of attributes being the last a binary classification.

3.3 Visualization and feature selection of DataSets

To analyze the datasets various kinds of data visualization, need to be taken into consideration to know the type of dataset the algorithm will encounter (e.g. correlation matrix, statistical data and data visualization graphs). The correlation matrix analysis or a scatter plot matrix will facilitate the understanding of the data and perform feature selection to improve the accuracy of the model.

3.3.1 Eye state Dataset

This Dataset was imported and analysed using a Seaborn heatmap to display the correlation between attributes. See Fig1.

As shown in Fig. 1. A few attributes have a high correlation, so it was decided to drop those with correlation>0.95. Resulting in a new dataset with 10 attributes instead of 14. With a maximun correlation of 0.60.

Fig. 1. HeatMap of Eye State DataSet.

3.3.2 Seizure Recognition Dataset

This Dataset is particularly big, (11500x178) so its manipulation can be at times difficult. Nonetheless, a correlation matrix with an implemented code to identify the high correlated attributes was useful to filter the dataset. The correlation had to be greater than 0.928 for the attribute to be dropped.

At the end, the Dataset was left with 7 attributes and the target. The Heatmap of the new dataset could then be visualized. See Fig. 2. Correlation in this case is helpful to have less redundant attributes and to enhance the accuracy of the method.

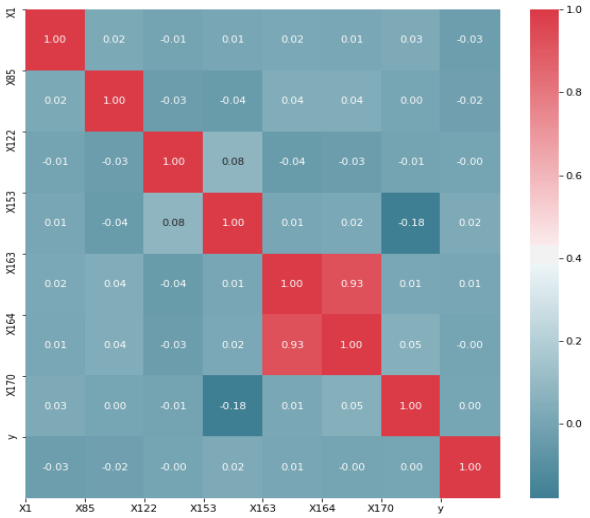

Fig. 2. HeatMap of Seizure Recognition Dataset.

3.3.3 Immunotherapy Dataset

This Dataset was analysed in three ways: with seaborn's factor plot being the hue the result of the treatment (Target), with a heatmap showing correlation (Fig. 8) and pairplot matrix (Fig. 7) showing the scatterplot among each one of the attributes.

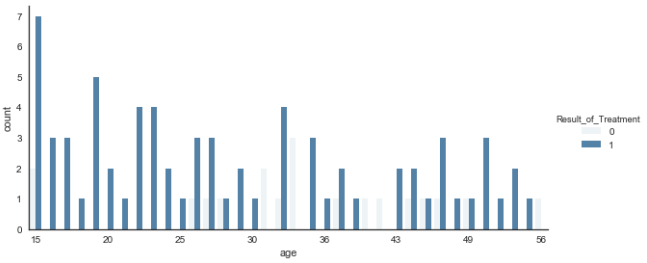

Fig. 3. Factor plot of age

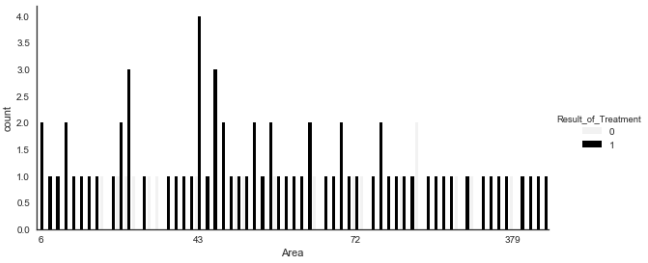

Fig.4. Factor plot Area

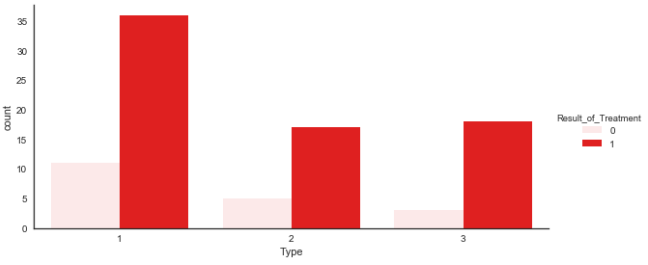

Fig.5. Factor Plot Type

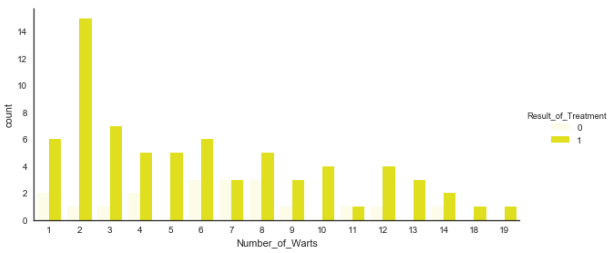

Fig. 6. Factor Plot of Number of Warts

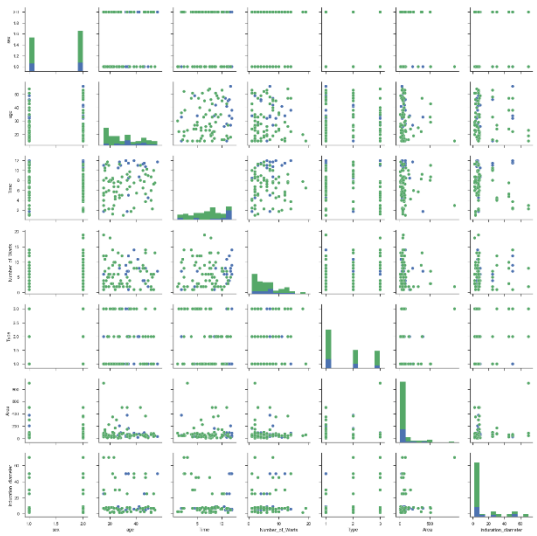

Fig. 7. PairPlot showing relationship between each pair of features

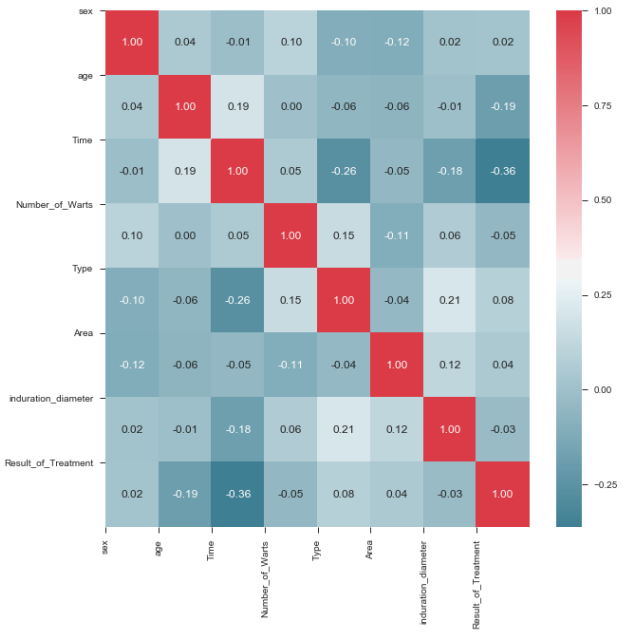

Fig. 8. HeatMap showing correlation factor.

As seen in Fig 8. This dataset has a low correlation among attributes which makes it unnecessary to drop attributes to enhance the performance.

4 Experiments

The overall objective of the project is to do "knowledge distillation" from a complex algorithm (i.e. Random Forest) to a simpler method (DT). A version of this strategy has already been pioneered by Rich Caruana and his collaborators [1]. Similarly, in [19] a neural network is used and obtains 99% of accuracy while classifying; later, a soft binary decision tree is used to learn from this previous classification and obtains 90.71% while a no-distillation-decision tree didn't perform as well, obtaining 80.88% of accuracy.

In this specific experiment the Scikit-learn toolbox in python will be used to obtain the probabilities of the RF using a function of the package that will output a matrix that contains the probabilities of each class per sample. After obtaining the probabilities a new dataset will be created making these past classifications the new target after discretizing the data. And a decision tree classifier from the same scikit-learn library will be used. The "visualize_classifier function" could be useful to have a better understanding of the classification process. This is the "student" function of the distillation process. The tree should be able to classify with a higher accuracy because of the "mimic" behaviour of the RF.

To do this, the probabilities are used to create different classes instead of just binary, after training the decision tree with these new classes, then the output should be binary again to be able to compare with the first dataset.

After training the decision tree with the probabilities of the RF, a new decision tree without the previous probability data will tested on the original dataset.

Also, a Neural Network (NN) and a Support Vector Machine (SVM) algorithm will be trained on the data set and will be put to predict. To later be compared with the distillation methodology.

To compare the performance of different Machine Learning Methods a table similar to Table 1. will be presented, as well as confusion matrix for each one of the predictions.

Table 1. Accuracy comparison example table.

DATASETS	ACCURACY				
	Random Forest	Distilled DT	Simple DT	SVM	NN

5 Discussion

The performances of the methods will be evaluated with a confusion matrix and tables that show accuracy information. Also, Tree visualization could also be useful to explain the classification of the data. In the same way, it might be useful to specify the feature importance of the classifiers to know which features are the most relevant for the determination of the classes; this can be done similarly with a function of the library in the RF.

To gain more insight about the performance of the methods, they will be compared with two other different methods. SVM and Neural Networks, will be trained with scikit-learn and their performance will be compared and contrasted with the previous.

7 Conclusion

The experiment should prove that distilling the knowledge from the RF and making a DT learn from its probabilities will enhance the performance of the DT in comparison with a Decision Tree without having a "teacher" model. Also, the use of a distilled-knowledge-DT should demonstrate that they can be a successful way of interpretation for cumbersome models. In the same way, the results obtained should be positive to use the model in healthcare projects where the interpretability is essential. It is expected that features importances indicated by the RF are comparable to the decisions made by the mimic DT.

The overall conclusion of the project should indicate and approve the methodology followed in this project of using these distilled models, while explaining the models' decisions. It should remark the use of interpretability and feature importance for health-related situations.

References

[1] G. Papamakarios, "Distilling Model Knowledge," 2015.
[2] N. Liberman, "Towards Data Science," 27 Jan 2017. [Online]. Available: https://towardsdatascience.com/decision-trees-and-random-forests-df0c3123f991. [Accessed 10 Feb 2019].
[3] S. Purushotham, Cornell Unversity, 11 Dec 2015. [Online]. Available: https://arxiv.org/abs/1512.03542. [Accessed 10 Feb 2019].
[4] G. Biau, "A random forest guided tour," TEST, vol. 25, no. 2, pp. 197-227, 2016.

[5] J. VanderPlas, "In-Depth: Decision Trees and Random Forests | Python Data Science Handbook," 2019. [Online]. Available: https://jakevdp.github.io/PythonDataScienceHandbook/05.08-random-forests.html. [Accessed 11 Feb 2019].
[6] C. Molnar, "4.4 Decision Tree | Interpretable Machine Learning," 2019. [Online]. Available: https://christophm.github.io/interpretable-ml-book/tree.html. [Accessed 11 Feb 2019].
[7] N. Donges, "The Random Forest Algorithm – Towards Data Science," 22 Feb 2018. [Online]. Available: https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd. [Accessed 11 02 2019].
[8] M. Olson, "Making Sense of Random Forest Probabilities: a Kernel," 2018.
[9] Y. Zhou, "Approximation Trees: Statistical Stability in Model," Ithaca NY, 2018.
[10] G. Cafri, L. Li, E. Paxton and J. Fan, "Predicting risk for adverse health events using random forest," Journal of Applied Statistics, vol. 45, no. 12, pp. 2279-2294, 2017.
[11] H. Kaur and D. Gupta, "Human Heart Disease Prediction System Using Random Forest Technique," International Journal of Computer Sciences and Engineering, vol. 6, no. 7, pp. 634-640, 2018.
[12] R. Díaz-Uriarte, "Gene selection and classification of microarray data using random forest," BMC Bioinformatics, vol. 7, no. 1, 2006.
[13] V. W. G.J.P, "Random Forest ensembles for detection and prediction of Alzheimer's disease with a good between-cohort robustness," NeuroImage: Clinical, vol. 6, pp. 115-125, 2014.
[14] D. K. Kumar, "Health Care Analysis Using Random Forest Algorithm," Journal of Chemical and Pharmaceutical Sciences.
[15] G. Hinton, O. Vinyals and D. Jeff , "Distilling the Knowledge in a Neural Network," 2015.
[16] "Machine Learning Repository," UCI, 2013. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/EEG+Eye+State#. [Accessed 10 Feb 2019].
[17] "Machine Learning Repocitory," UCI, 24 May 2017. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Epileptic+Seizure+Recognition. [Accessed 10 Feb 2019].
[18] Q. Wu, "Machine Learning Repository," UCI, [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Epileptic+Seizure+Recognition. [Accessed 10 Feb 2019].
[19] lmartak, "Github," 2017. [Online]. Available: https://github.com/lmartak/distill-nn-tree/blob/master/mnist.ipynb. [Accessed 10 02 2019].