Project 3: Knowledge distillation from random forests (Supervised learning)

Student ID 1806019

***Abstract*—A fundamental strategy to improve the execution of all machine learning algorithms is to train different models on comparative data and then to average their result to get better predictions. If the individual models are extensive neural networks it is cumbersome to make predictions using a whole group of models and can be extremely computationally expensive to send to users. Knowledge Distillation releases and injects the data in a complex model into a simpler model. A general framework is introduced for knowledge distillation then a supportive model dependent on our own preference makes sense of how to imitate a complicated model. In this paper I will show the accuracy that the random forest trained on the data sets has acquired. Random forest builds multiple decision trees and combines them together to make a prediction more accurate and stable, after this the decision tree algorithm will be trained by the probabilities calculated by the random forest and lastly accuracies generated by the two classifiers will be compared.**

I. Introduction

Late progress in machine learning was exceptional. Record after record is broken as models are gradually becoming more accurate and performing better [1]. Progression of theoretical and algorithmic achievements, mechanical advancements in PC equipment and programming made it possible. Deep neural networks can be moderate to evaluate and costly to store with a few concealed layers.

Imagine approaching a precise and innovative model that is stubborn and badly intended to be used anyway but can be replaced with a helpful model that works as well and is just as great as it is. This method is called Knowledge Distillation; as information can be extracted from complex model, refined it and saturated it into the favourable model. If the beneficial model fails to effectively imitate it complex model, it will be ignored. In Knowledge distillation, the information is transferred to the developed model by preparing it on an exchange set and using a soft target distribution for each case in the exchange set delivered by using the complex model with high softmax temperature.

In this paper I shall describe that how random forest will be used to calculate probabilities, how knowledge distillation will be done and how accuracies of new and original data set will be compared. A Random Forest, which simplifies the procedure, is basically an ensemble of decision trees. Simply put, random forest builds and merges multiple decision trees to achieve a more accurate and stable prediction. The method of bagging trains random forests. For those algorithms that have high variance, bagging is a general procedure that can be used to reduce the variance. A highly variable algorithm are decision trees, such as classification and regression trees. This technique allows repeated use of a few occurrences for the training phase as we are testing with replacement. By applying the bagging technique to the element space, the Random Forest strategy presents more arbitrariness and wide range[2]. Upon training the random forest, the probabilities are now obtained for each class and the probabilities that are calculated by the random forest.

Then the new dataset will be created and the distillation of knowledge is done, knowledge shall be shared into the more refined model. It is done by preparing it on an exchange set and soft target distribution is utilized for each case in the exchange set which is created by complex model with a high softmax temperature. In preparing the refined model a similar high temperature is used but it uses a temperature of 1. If all or part of the exchange set known on the right labels, this approach can basically be improved by preparation of a refined model to supply the right labels. The main objective is the log loss of the soft target and it is recorded using the same high temperature as was used to produce the soft targets from the complex model, and find the right labels[3]. Numerical variables will be transformed in categorical counterparts and numpy functions will be used to create probability bins. Decision tree classifiers will be learned on the new dataset, plotted and measured on the original datasets for their accuracy.

II. Background

Knowledge distillation with neural networks was pioneered by [4], a transfer learning method aimed at enhancing the training of a network of students by relying on knowledge borrowed from a powerful teacher network. Although in some special cases it has been shown that shallow networks can approximate deeper networks without loss of accuracy [5] later work related to knowledge distillation was mostly based on the assumption that deeper networks are always learning better representations. For example, by using a shallow one with more parameters, [6] tried to learn a thin deep network. The introduction of highway [7] and later residual networks enabled highly accurate training of very deep architectures, and the generality of these networks was demonstrated experimentally across a wide variety of datasets. Distillation of knowledge is a principled approach to accelerating small neural networks.

An extensive list of papers tried to transfer knowledge for different purposes between one model and another. The goal is sometimes compression: to produce a compact model that preserves the accuracy of a larger model that takes up more space and requires more prediction computing [8] proposed that neural networks and multipletree predictors be compressed by a single tree approximation. More recently, others suggested transferring knowledge from neural networks to simpler models such as decision trees and generalized additive models [9] to increase interpretability.

Furthermore, for the purpose of explaining decisions, [10] proposed to distil deep networks into decision trees. For all the researchers, the goal of knowledge transfer is to produce a student model that achieves better accuracy by transferring knowledge from the teacher model as if it were directly trained. The resource constraints of underpowered devices such as cellphones and internet of things devices often motivate this research. [3] compresses the information in a neural network ensemble into a single neural network in a pioneering work. With modern deep learning tools, [4]

demonstrated a method to increase the accuracy of shallow neural networks by training them to imitate deep neural networks by penalizing the difference between the student's and the teacher's logits in the L2 standard.

[11] forces the student at the end of each residual stage to match the teacher's attention map. In addition to minimizing the divergence from teacher predictions, [12] seek to minimize the difference between teacher and student loss derivatives with respect to input aimed to compress models by approximating the mappings between hidden layers of teacher and student by using linear layers of projection to train relatively smaller students, who exhibited a method called dark knowledge in which a student model trains with the objective of matching the teacher model's full softmax distribution, which increased interest in Knowledge Distillation. One publication that applied ML to Higgs Boson and detection of supersymmetry made the leap to apply dark knowledge to the search for dark matter [5].

For example, [9] distill multiple DQN models into one in the deep reinforcement learning community. Several recent papers [8] use KD to minimize oblivion in ongoing learning and include it in the training scheme for adversaries. [5] have recently highlighted some connections between KD and a privileged information learning theory. proposed applying KD from a DNN to another DNN of the same architecture, reporting that the student model trains more quickly and attains greater accuracy than the teacher.

III. Methodology

A concise overview of the data sets will help to provide a clear overview of the different attributes present in the data. Jupyter notebook will be used as the development environment for this system, with the first step being to import the required libraries such as pandas, numpy and scikit learning environment. The next step is to load the dataset after importing the libraries. In actual-world problems, data is never tidy and finish, partly, mostly inconsistent, does not provide the correct outlook, includes outliers, noise, missing values, lack of attribute values, irregular patterns,lack of certain attributes of interest or involving only Overall data, involving errors, actually contains discrepancies in codes or names. To tackle this problem, the best solution is to perform a set of operations to clean, massage and arrange the raw data in an entirely justified form, this set of operations or processes is called "Preprocessing the data." Head function will be used to test quickly if object has the right type of data in it after importing the dataset. Next step is to check the Null values. If the missing values are not properly handled then we may actually draw an inaccurate data inference. Using the mean imputation technique, missing values will be imputed if any column has more than 70% missing values then the column will be dropped. This approach can only be done on numeric values if the data is categorical then it will be replaced by numeric values and the LabelEncoder) (class from the preprocessing library will be used to convert categorical values into numeric values. Data outliers will also be detected and addressed. Some machine learning require scaling, standardization and multicolinearity to check the dependence of features on the data set To properly model the data. Validation set approach can be used to split and sample data, dividing data into test set and train set, dividing data into 70:30, 80:20 or 50:50 split ratios to train the model. Random Forest algorithm will also be trained using train data, predictive accuracy along with accurate recall will be calculated and F1 scores will be calculated once the model is trained. Model prob function will be used to generate the target variable probabilities. A probability bin is generated that contains a distribution of prediction intervals, the model's generated probabilities are redistributed according to the present distribution.

Based on each target variable's class. The induction of the probabilities generated after binning will make a new dataset. This set of data will be categorized at a multi-class level and then the distillation of knowledge will be carried out. The new dataset will be trained to predict their accuracy before and after the distillation procedure on the decision tree algorithm. To determine the level of success, this performance is also compared to other classification algorithms.

Let's see some of the terms mentioned here

A. Random Forest

RF is a Regression and Classification Trees collection or ensemble trained on data sets with the same size as the training set, named bootstraps, generated from a random re sampling of the training data set itself. A collection of bootstraps, which does not include any specific record from the initial[out - of-bag (OOB) samples] dataset, is used as a test set once a tree is constructed. Breiman (1996) Demonstrated by empirical evidence that the OOB error is accurate for the bagged classifiers to use a test set with the same size as the training set. Using the estimate of OOB therefore eliminates the use of a separate test set. Every single CART tree votes for one class to identify new input data, as well as the forest predicts the class that gets the majority of the vote.

RF follows special rules for tree growth, self-testing ,post-processing and tree combination, is robust to overfit and is taken into consideration more stable in the existence of outliers and in very large parameter spaces than other machine learning algorithms [3].

The principle of variable importance is an implicit selection of features performed by RF using a random subspace methodology and is evaluated by the Gini. criterion impurity index .

The concept of variable importance is an implicit feature selection performed by RF with a random subspace methodology, and it is assessed by the Gini impurity criterion index. The

Gini index is a predictive power measure of classification or regression variables based on the impurity reduction principle (Strobl et al., 2007); it is anti-parametric and does not rely on data depending to a particular type of distribution. The Gini index of a node n is calculated as follows for a binary split.

Gini( )=1−∑2j=1(j)2Gini(n)=1-∑j=12(pj)2

Where pj is class j's relative frequency in node n.

The enhancement in the Gini index should be increased to divide a binary node in the best possible way. In other statements, a low Gini (i.e., a greater decrease in Gini) implies that a specific predictor feature performs a important role in dividing the data into the two classes. The Gini index can therefore be used to rank features important for a classification issue.

B. Knowledge Distillation

The word "knowledge distillation" came from Hinton et al. (2015)'s recent work. To the best of our knowledge, Caruana et al. first proposed the use of knowledge transfer (KT) to compress model[50]. They trained a compressed / ensemble model of strong classifiers with the labeling of pseudo-data and reproduced the original larger network output. But the work is limited to models that are shallow. Recently, the idea was introduced as knowledge distillation (KD) to compress wide and deep networks into shallower ones, where the compressed model imitated the function learned by the complex model. The key idea of KD-based methods is to transfer knowledge from a large teacher model to a small one by using softmax to learn the output of the class distributions. The work in introduced a framework for KD compression, which eased deep network training by following a student-teacher paradigm in which the student was penalized based on the softened version of the teacher's output. The structure compressed a teacher network ensemble into a similar-depth student network. In order to predict output and classification labels, the student was trained. KD shows promising results in various image classification tasks despite its simplicity. The work aimed at addressing the problem of network compression through the use of neural networks. It suggested an approach for compressing wide and shallower (but still deep) networks to train thin but deep networks called FitNets. The method has been extended to allow student models to be thinner and deeper. To learn from the teacher network's intermediate representations, FitNet made the student mimic the teacher's full feature maps. Such assumptions, however, are too strict as teacher and student capabilities may vary greatly.

There are several enhancements of distillation knowledge along this direction. The work in has trained a model of parametric students to approximate a teacher from Monte Carlo. The proposed structure utilized online training and the student model used deep neural networks. In contrast to previous works representing knowledge using the soft label probabilities,the knowledge was represented by using the neurons in the higher hidden layer, which maintained as much data as the probabilities of the label but are more compact. The research in accelerated the experimentation method by immediately transferring the knowledge to each new deeper or wider network from a previous network. The strategies are based on the neural network specifications concept of function-preserving transformations. Attention Transfer (AT) proposed by Zagoruyko et al.[57] to relax FitNet's assumption. They transmitted the attention maps summarizing the complete activations.

Drawbacks: KD-based approaches can thin out deeper models and help reduce computational costs significantly. There are some disadvantages, though. One of those is that KD could only be applied with the softmax loss function to classification tasks, which hinders its use. Another drawback is that the model assumptions are sometimes too strict to make the performance competitive with other approaches.

C. Decision Tree:

Decision Trees (DTs) is a non-parametric supervised method of learning used to classify and regress. Decision trees learn to approximate a sine curve with a set of if-then-else decision rules from the data. The deeper the tree, the more complex the rules of decision and the fitter the model. Decision tree constructs a tree structure in the form of regression or classification models. It splits down a set of data into smaller and smaller subsets and at the same time increasingly developing an associated decision tree. The end result is a tree with nodes of decision and nodes of leaf. There are two or more branches in a decision node. Leaf node is a classification or decision. The top decision node in a tree that matches the best predictor called root node. Both categorical and numerical data can be handled by decision trees.

D. Data Description:

This paper includes the three datasets from the UCI repository and Kaggle were used to evaluate the Knowledge distillation performance along with the classification models. The dataset 1 bank direct marketing related to a Portuguese banking institution's direct marketing campaigns. The campaigns for marketing were based on phone calls. Often, in order to access the product(bank term deposit), more than one contact with the same customer was required. The banl direct marketing data set contains (45211) number of samples without missing values with 17 attributes. Data set features consisting of two types: nominal and numeral attributes, as shown in Figure-1. This table shows that three types of attributes are categorical as the attributes (Job, Education, Contact, Marital, Month, Poutcome) in the range type of all attributes (Age, Day duration, Balance, Campaign, Days and Previous). And binary categories are all attributes represented in their classes as Yes or No: attributes (Default, Housing, Loan, Output) for example.

Figure-2 holds the data which is related to the rain were based on the Australia, The target task is to predict if the rain will come tomorrow or not? It is a classification problem, The model will give yes of no two classes. Target Variable name is RainTomorrow. This dataset has 145460 observations and 24 variables.

Figure-3 holds the The data which is related to the heart disease in the patient.The target task is to predict if the client

Variables	Description
Age	Customer's Age
Job	type of job
Marital	marital status
Education	Csutomer's Education
Default:	has credit in default?
Housing	has housing loan?
Loan	has personal loan?
Contact	contact communication type
Month	last contact month of year
Day_of_week	last contact day of the week
Duration	last contact duration
Campaign	campaign and for this client
Pdays	number of days that passed by after the client was last contacted from a previous campaign
Previous	number of contacts performed before this campaign and for this client
Poutcome	outcome of the previous marketing campaign
Emp.var.rate	employment variation rate - quarterly indicator
Cons.price.idx:	consumer price index - monthly indicator
Cons.conf.idx	consumer confidence index - monthly indicator
Euribor3m	euribor 3 month rate - daily indicator
Nr.employed	number of employees - quarterly indicator
y	has the client subscribed a term deposit?

Fig. 1. Data about Bank marketing

Variables	Description
Date	The date of observation
Location	The common name of the location of the weather station
MinTemp	The minimum temperature in degrees celsius
MaxTemp	The maximum temperature in degrees celsius
Rainfall	The amount of rainfall recorded for the day in mm
Evaporation	The so-called Class A pan evaporation (mm) in the 24 hours to 9am
Sunshine	The number of hours of bright sunshine in the day.
WindGustDir	The direction of the strongest wind gust in the 24 hours to midnight
WindGustSpee	The speed (km/h) of the strongest wind gust in the 24 hours to midnight
WindDir9am	Direction of the wind at 9am
WindDir3pm	Direction of the wind at 3pm
WindSpeed9an	Wind speed (km/hr) averaged over 10 minutes prior to 9am
WindSpeed3pn	Wind speed (km/hr) averaged over 10 minutes prior to 3pm
Humidity9am	Humidity (percent) at 9am
Humidity3pm	Humidity (percent) at 3pm
Pressure9am	Atmospheric pressure (hpa) reduced to mean sea level at 9am
Pressure3pm	Atmospheric pressure (hpa) reduced to mean sea level at 3pm
Cloud9am	Fraction of sky obscured by cloud at 9am. This is measured in "oktas", which are a unit of eigths. It records how many eigths of the sky are obscured by cloud. A 0 measure indicates completely clear sky whilst an 8 indicates that it is completely overcast.
Cloud3pm	Fraction of sky obscured by cloud (in "oktas": eighths) at 3pm. See Cload9am for a description of the values
Temp9am	Temperature (degrees C) at 9am
Temp3pm	Temperature (degrees C) at 3pm
RainToday	Boolean: 1 if precipitation (mm) in the 24 hours to 9am exceeds 1mm, otherwise 0
RISK_MM	The amount of next day rain in mm. Used to create response variable RainTomorrow. A kind of measure of the "risk".
RainTomorrow	The target variable. Did it rain tomorrow?

Fig. 2. Data about Rain

will get heart attack or not. This dataset has 303 observations and 14 variablesSo, I did some feature engineering to clean the data for model fitting.

Variables	Description
age	age in years
sex	(1 = male; 0 = female)
cp	chest pain type
trestbps	resting blood pressure (in mm Hg on admission to the hospital)
chol	serum cholestoral in mg/dl
fbs	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
restecg	resting electrocardiographic results
thalach	maximum heart rate achieved
exang	exercise induced angina (1 = yes; 0 = no)
oldpeak	ST depression induced by exercise relative to rest
slope	the slope of the peak exercise ST segment
ca	number of major vessels (0-3) colored by flourosopy
thal	3 = normal; 6 = fixed defect; 7 = reversable defect
target	1 or 0

Fig. 3. Data About Heart

IV. Experiments

For this research I have chosen three data sets on the basis of a classification task. Every target variable in the dataset depends to a certain class. A similar pattern follows the procedure followed in the experiment for all three datasets. All features within the dataset are standardized so that a Gaussian distribution follows the distribution among the variables. A random forest with a reasonably high number of trees is trained on the datasets. Within the dataset, the probabilities of each class are calculated and a new dataset containing these probabilities is included. Decision tree classifiers (Distillation model) are then trained on the new datasets and compared to the original data set for their performance. Many state-of - the-art machine learning algorithms such as logistic regression, SVM, KNN, Naive Bayes are trained on both the original and new datasets as well as comparing performance metrics.

These three datasets where different from each other. So, checked missing values in all data sets, some values replaces with column mean value but certain records we could not impute, if that column has categorical values, its not possible to replace the mean values.

A. Bank marketing Dataset:

As already mentioned about the dataset in the above topic, so here explaining about the pre-processing and experiment for all the three datas. bank dataset. So, I did some feature engineering to clean the data for model fitting. In this, most of the variables converted from categorical to numeric to do the classification task.

Figure-4 explains about after feature engineering how data looks like for model fit. All variables converted into numeric and continuous variables converted.

Figure-1 shows the distribution of credit default, Housing loan, Loan based on the customer. We can see around 10000 of customer has credit issues. Next, more than 20000 customer have taken home loan, following by very less people have taken the loans.

	age	job	marital	education	default	housing	loan	contact	month	day_of_week	...	emp.var.rate	cons.price.idx
0	3	3	1	0	0	0	0	1	6	1	...	1.10	93.99
1	3	7	1	3	1	0	0	1	6	1	...	1.10	93.99
2	2	7	1	3	0	2	0	1	6	1	...	1.10	93.99
3	2	0	1	1	0	0	0	1	6	1	...	1.10	93.99
4	3	7	1	3	0	0	2	1	6	1	...	1.10	93.99

5 rows × 21 columns

Fig. 4. Final features for model-Bank marketing

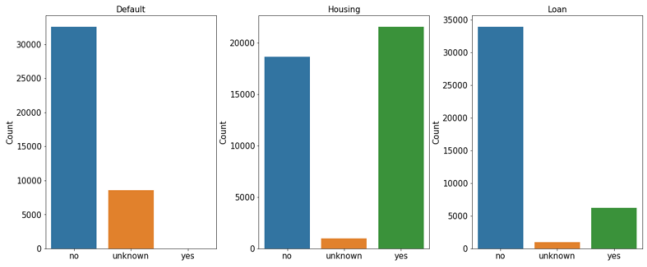

Fig. 5. Credit Default,Housing Loan, Other Loan

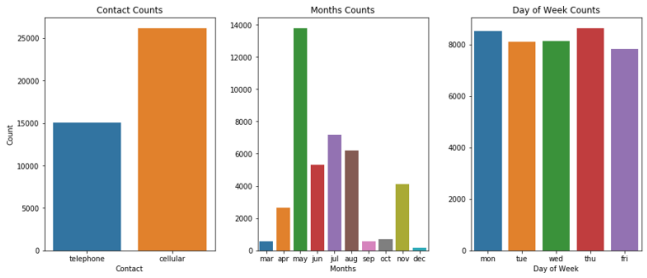

Fig. 6. Contact Information, Month& Date of Contact

Figure-2 shows ,the data of Contact information , month and Days count of the customer. Because, to do the marketing call, they should be aware of this data. By seeing first bar chart, mostly they can call to the customer's cellular, and maximum calls done on may month, as well as most calls happened on Monday and Thursday.

Now let's move on to the next part of the study in this, where modeling will take place. After cleaning all the features within the dataset, there will be a sample split where the data set will be split to train and test according to the 50:50 ratio. Data will be trained on a random forest model with a number of trees equal to 150 and their probabilities will be calculated

By using binning we have converted the class classification to multi-class classification. Then applied the decision tree model against it the new dataset which contains multiclass probabilities after distillation. When running the RF into the original Dataset and it prediction accuracy was 90 percent. After distilled model, the prediction of accuracy has increased by a 5% with score of 95%.

B. Rain Data Set:

In previous section discussed about the Data. So, I did some feature engineering to clean the data for model fitting. This data is quite decent tiny data. So, little bit feature selection done with this data. In this, most of the variables converted from categorical to numeric to do the classification task.

	MinTemp	MaxTemp	Rainfall	WindGustSpeed	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity3pm	Pressure9am
4	0.628342	0.696296	0.035714	0.465753	0.135135	0.428571	0.797753	0.33	0.342043
5	0.550802	0.632099	0.007143	0.671233	0.459459	0.523810	0.494382	0.23	0.304038
6	0.542781	0.516049	0.000000	0.589041	0.486486	0.523810	0.426966	0.19	0.313539
7	0.366310	0.558025	0.000000	0.383562	0.108108	0.357143	0.415730	0.19	0.403800
8	0.419786	0.686420	0.000000	1.000000	0.135135	0.619048	0.348315	0.09	0.296912
9	0.510695	0.641975	0.050000	0.287671	0.351351	0.214286	0.528090	0.27	0.251781

Fig. 7. Final Features for Model Fit

Figure-7 shows the features ready for model fit.

Now let's move on to the next part of the study in this, where modeling will take place. After cleaning all the features within the dataset, there will be a sample split where the data set will be split to train and test according to the 30:70 ratio. Data will be trained on a random forest model with a number of trees equal to 100 and their probabilities will be calculated

By using binning we have converted the class classification to multi-class classification. Then applied the decision tree model against it the new dataset which contains multiclass probabilities after distillation. When running the RF into the original Dataset and it prediction accuracy was 85 percent. After distilled model, the prediction of accuracy has increased by a 2% with score of 95.

C. Heart Disease Data Set

I did some feature engineering to clean the data for model fitting. In this, most of the variables converted from categorical to numeric to do the classification task. After feature selection,

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Fig. 8. Final Features for Heart Disease

Figure-8 Shows that after feature Engineering this above features ready for model fit.

Figure-9 Explains the co-relation between the variables. Now let's move on to the next part of the study in this, where modeling will take place. After cleaning all the features within the dataset, there will be a sample split where the data set will be split to train and test according to the 20:80 ratio. Data will be trained on a random forest model with a number of trees equal to 100 and their probabilities will be calculated by using binning we have converted the class classification to multi-class classification. Then applied the decision tree model against it the new dataset which contains multiclass probabilities after distillation. When running the RF into the original Dataset and it prediction accuracy was 81 percent. After distilled model, the prediction of accuracy has increased by 12% with score of 93%.

Statistical measures are used to evaluate the performance of each classification model; classification accuracy, sensitivity, and specificity. Kohavi and Provost, 1998, define these

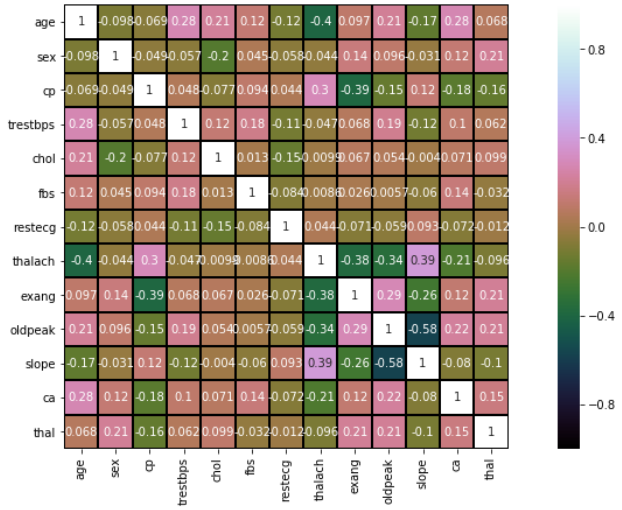

Fig. 9. Correlation

TABLE I
STAGES OF EXPERIMENTS

Steps	Process
1	Pre-Processing
2	Spliting the Datas
3	Random Forest to predict probability
4	Binning that probability
5	Convert Binary classification into multi-class classification
6	Decision Tree model with the new data
7	Other Algorithmns with the original Data
8	Other Algorithmns with the New Data
9	Comparing Accuracy

measures as a confusion matrix, containing information on actual and predicted classifications performed by a classification system. It uses true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). Correct / Incorrect classification percentage is the difference between the variables ' actual and predicted values. True Positive (TP) is the number of correct predictions, or in other words, that an instance is true; This occurs when a positive classifier prediction coincides with a positive target attribute prediction. True Negative (TN) presents a number of correct predictions that an instance is false, (i.e.) when both the classifier occurs, and the target attribute suggests that there is no positive prediction. The number of incorrect predictions that an instance is true is the False Positive (FP). Finally, the number of incorrect predictions that an instance is false is False Negative (FN). Table 4 shows a two-class classifier's confusion matrix.

Accuracy of classification is defined as the ratio of the number of cases properly classified and is equal to the sum of

		Predicted	
		Positive(yes)	Negative(no)
Actual	Positive(Yes)	TP	FP
	Negative(no)	TN	FN

Fig. 10. Confusion Matrix

TP and TN divided by the total number of cases N[21].

$$Accuracy = \frac{TP + TN}{N}$$

Sensitivity pertains to the rate of positive properly classified and is equal to TP divided by the sum of TP and FN. Sensitivity can be called a True Positive Rate.

$$Sensitivity = \frac{TP}{TP + FN}$$

Specificity corresponds to the properly classified negative rate and is equal to the TN-to-TN-to-FP ratio[21].

$$Specificty = \frac{TN}{TN + FP}$$

Model evaluation can be performed using multiple metrics. Here, While seeing the Table 2, All algorithms performing well after knowledge transfer done to the new data set. DT accuracy was 88% but after distillation it rose to 95%. Compare to all algorithms Navie bayes Algorithm got less score, but it performed well in the new data set.

TABLE II
BANK DATASET ALGORITHMS ACCURACY

Algorithms	Original Dataset	New Dataset
DT	88	95
LR	90	95
KNN	89	91
SVM	90	93
NB	84	92

As Discussed earlier in the second data set performed well after distillation. Moreover, Naive Bayes Algorithm performing less when compare to other algorithms.

TABLE III
ACCURACY OF RAIN DATASET WITH DIFFERENT ALGORITHMS

Algorithms	Original Dataset	New Dataset
DT	82	95
LR	84	85
KNN	82	84
SVM	82	84
NB	79	83

In this Data set, SVM performed less when compare to other algorithms, but it was significant increase when applied knowledge distillation concept.

Moreover, Its depends on the data set quality we can get the accuracy level.Knowledge distillation working well with all the data, but I tried in the large data but it gives slightly improvement to the model.The enhancement shown by the decision tree is due to the information obtained with

TABLE IV
ACCURACY OF HEART DATASET WITH DIFFERENT ALGORITHMS

Algorithms	Original Dataset	New Dataset
DT	71	93
LR	81	96
KNN	65	67
SVM	56	57
NB	83	96

the original dataset by Random Forest.Probability passing is just one of many ways to distill knowledge. If we examine each table, each table shows percentage improvement when knowledge distillation is performed and also the performance tends to increase or show improvement when features of the previous model are included in the dataset. Current knowledge distillation methods needs full training data to distill knowledge from a vast network of "teachers" to a spacious network of "students" by combining those statistics between "teacher" and "student" such as softmax outputs and feature responses[13]. Not only is this time-consuming, and also contradictory with human behavior where, with few examples, children can learn knowledge from adults[13]. A table must have two rows and two columns[2] for a binary classification problem in confusion matrix.

Other way to validate the accuracy level to check the error rate. below given error rate for each table,which means after distillation our model performed well with the new dataset.

```
Mean Absolute Error: 0.046856033017722745
Mean Squared Error: 0.046856033017722745
Root Mean Squared Error: 0.2164625441449923
```

Fig. 11. Errors in Data set 1

```
Mean Absolute Error: 0.04236581069806248
Mean Squared Error: 0.04236581069806248
Root Mean Squared Error: 0.20582956711333403
```

Fig. 12. Errors in Data set 2

```
Mean Absolute Error: 0.06557377049180328
Mean Squared Error: 0.06557377049180328
Root Mean Squared Error: 0.25607375986579195
```

Fig. 13. Errors in Data set 3

V. CONCLUSION

In all the dataset used in this study knowledge distillation has improved the accuracy siginificantly. Although in smaller datasets it looks more efficient as compared to the the larger datasets. Moreover, random forest is one of the efficient algorithm in the current machine learning field, when compared to other algorithms. Knowledge distillation is one of the newest concept for model compression, but it has proved to be one of the important method to improve model accuracy.

Furthermore, selection of features or extraction technique such as PCA is not used to select features as random forests have the tendency to select the best modeling features. In the future studies this technique can be used on platforms where running the cumbersome model is not possible. In near future, with more advancement in the knowledge distillation and related fields, we may able to use the refined and simplest form of cumbersome model for our daily purposes as well to predict the occurrence of an event by feeding the data.

VI. REFERENCES

[1] I-Cheng Yeh and Che-hui Lien. “The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients”. In: Expert Systems with Applications 36.2 (2009), pp. 2473– 2480.

[2]Orgler,Y.E.1971EvaluationofBankConsumer Loans with Credit Scoring Models. Journal of Bank Research 2 (1): 31-37

[3] Lee, T., Chiu, C. Lu, C., Chen, I. 2002. Credit Scoring Using the Hybrid Neural Discriminant Technique. Expert Systems with Applications 23 (3): 245-254

[4] Emel, A., Oral, M., Reisman, A., Yolalan, R. 2003. A credit scoring approach for the commercial banking sector. Socio-Economic Planning Sciences 37 (2): 103-123.

[5] Hand, D. J., Mannila, H., & Smyth, P. (2001). Data mining: Practical machine learning tools and techniques. Cambridge: MIT Press.

[6]Xuan Liu, Xiaoguang Wang, and Stan Matwin. “Improving the Interpretability of Deep Neural Networks with Knowledge Distillation”. In: 2018 IEEE International Conference on Data Mining Workshops (ICDMW). IEEE. 2018, pp. 905–912.

[7] Sustersic, M., Mramor, D., Zupan J. 2009. Consumer credit scoring models with limited data. Expert Systems with Applications 36 (3): 4736-4744.

[8] ”Detecting Fraud in Financial Payments”STANFORD UNIVERSITY in collaboration with Maikel Lobbezoo *December 15, 2017.

[9] ”Detection of financial statement fraud and feature selection using data mining techniques”,P. Ravisankar,V. Ravi, G. Raghava Rao ,I. Bose.(2011) 491500.

[10] ”The alternating decision tree learning algorithm” Yoav Freund, Llew Mason.

[11] ”Rotation Forest:A New Classifier Ensemble Method Juan J. Rodrguez, Member, IEEE Computer Society, Ludmila I. Kuncheva, Member, IEEE, and Carlos J. Alons,October 2006.

[12] Data Mining Framework for Direct Marketing:A Case Study of Bank Marketing,Lilian Sing’oei1 and Jiayang Wang2,School of Information Science and Engineering, Central South University
Changsha, 410083, China

[13] R. Nisbet, J. Elder and G. Miner. Handbook of statistical
analysis and data mining applications. Academic Press, Burlington, MA, 2009.[1]

[1]Githhub Id-varadharajanviji/**CE888lab**