# Distillation using Random Forest Classifier on Decision Tree Classifier with Comparison

***Abstract: In the paper a descriptive analysis of three binary classification dataset and one continuous dataset, along with train and testing of the machine learning model are explained and compared. The machine learning model are Random Forest Classifier, Decision Tree Classifier with distillation knowledge and without distillation knowledge are trained and tested. Distillation is performed on the target variable using Random Forest Classifier. The comparison is made using accuracy and performance of the model. The comparative studies clearly explain the distillation knowledge on target variable improve performance of the model if the target variable is in continuous form.***

1. **INTRODUCTION**

Machine learning (ML) model plays a major role in today’s world. It automates the model/machine, learn from the past data and make decisions based on training model. If the amount of the dataset is large it helps to build models more accurate but sometimes it results into overfitting problem or complex of the model increases, results into more time consuming and incorrect prediction.

Thus, to overcome the problem of large scale dataset, the dataset is divided into groups and using small scale dataset we can obtain same learning and process done using large scale dataset but without complex training and overfitting problem using Distillation knowledge.

* 1. **Overfitting:** When the model is so precisely trained that it negatively performs on the test dataset, which results into less accurate model and prediction, it is termed as overfitting. As, it learn noise or outliers of the training data too along with the dataset.
  2. **Distillation:** Distillation is a process, where the training dataset is compressed into bins and used by the model to train and test, which not only reduce the processing time but the result are almost same without distillation.

1. **LITERATURE REVIEW**

For knowledge distillation, there are mainly two points to be considered, first extract all knowledge from the dataset and second pass the knowledge so it can be used further. Thus, it may face lots of challenges, hence a comparative studies and performance of the distillation knowledge is measured in [6], which shows Rocket knowledge distillation accuracy is improved the highest by 1.25%.

Distillation knowledge is used to reduce the complexity of the model, hence used in deep learning model for computer vision tasks [7] and performance of the model, distillation architecture, and applications are measured to get better insight of the process.

1. **METHODOLGY**
   1. **Dataset:** For the given task, we have preferred four datasets from UCI repository [1]. The first dataset is about the person earning over 50K in a year or not based on the sex, age, occupation types of feature – Adult Data Set [2]. Second dataset is to build a classification model to predict whether the firm is fake or not based on the risk factors for the auditors – Audit Dataset [3]. Third dataset gives a real result of treatment from the Immunotherapy – Immunotherapy Dataset [4]. All the above dataset contains binary target variable and hence used to train and test binary classification model. Below are the images for few of the samples of the three dataset.

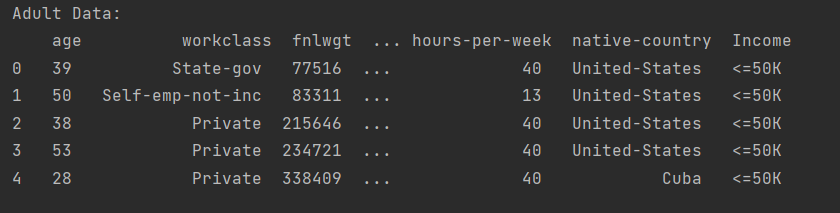
****

Figure Adult Dataset

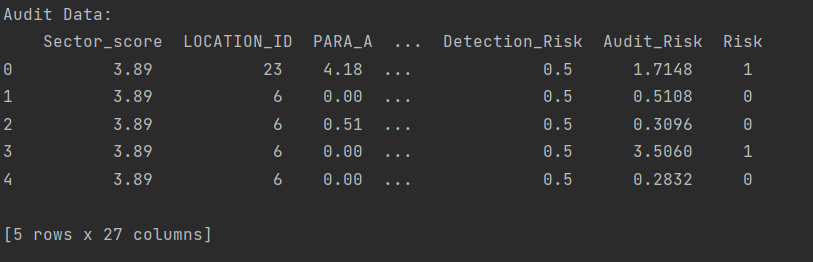


Figure Audit Dataset

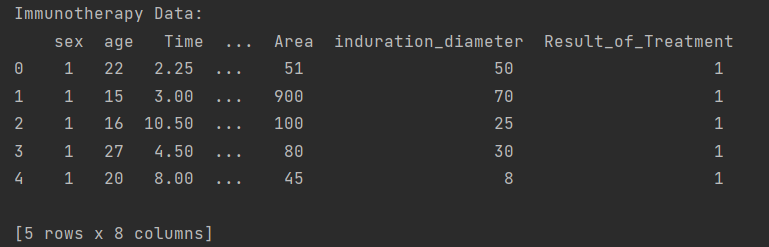
******

Figure Immunotherapy Dataset

The fourth dataset contains target variable in a continuous form, about the acres of land destroyed in a forest fire in a northeastern region of Portugal – Forest Fires Dataset [5].

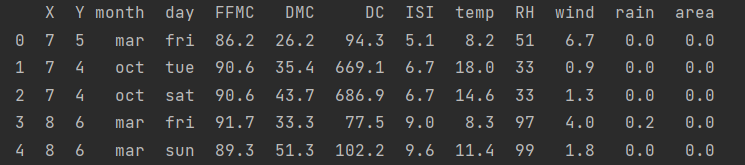


Figure Forest Fires Dataset

* 1. **Brief Explanation of Different Steps:** The models used in the assignments are Random Forest Classifier and Decision Tree Classifier. Below are the steps take to perform the task.

1. Required python libraries are imported to perform the required task – seaborn, matplotlib, pandas, sklearn.
2. Pre-processing of the Data and Data Exploration:
3. Using pandas the dataset file is read and converted into dataframe for further processing.
4. With the help of pandas inbuilt function ‘dtypes’, ‘head()’, and ‘describe()’, the dataset is explored and the column datatype, sample of data and description of the dataset is presented, which illustrates the data into depth.
5. A label encoder is used for textual data in the dataset and converts all the textual data into numeric data. Label encoder converts into unique numeric value for a unique textual data. For example, sex – male and female would be converted into 1 and 0 using label encoder. Any descriptive texts in the dataset are converted using TF-IDF Vectorizer.
6. Using seaborn and matplotlib, the bar plot is plotted.
7. Machine Learning Model:
   1. The target variable is separated from the feature variable for splitting the dataset.
   2. The dataset is split with 70% of the dataset used in training of the model, while remaining 30% dataset are used for testing of the model.
   3. A Random Forest Classifier object is called with number of tree equal to 500.
   4. The object is trained and tested using the training and testing dataset.
   5. Using accuracy of the classification model and mean squared error (MSE) for regression model is taken into consideration for the performance of the model.
   6. Using predicting probability the bins are created for the target variable and stored into two different columns in the dataframe of the dataset, which is further used for the target variable in the decision tree model as distillation knowledge.
   7. Now the target values are multiclass and multilabel in form.
   8. Using Grid Search CV, best hyperparameters of the decision tree model are derived for the given dataset.
   9. Now the distillation knowledge target variable along with other feature variable is splitted with 70-30% ratio.
   10. The training of the model of the decision tree model is performed and tested based on the derived distillation variable and other feature variable.
   11. To compare the result of the distillation process, the decision tree model is trained again but with the original dataset and tested.
   12. Decision Tree Graph is saved for each dataset in distillation process and without distillation process is visualized.
8. **RESULTS**
   1. **Random Forest Classifier/Regression:** Below table shows the accuracy of the model achieved for the given three binary classification dataset and MSE for the continuous target dataset.

|  |  |  |  |
| --- | --- | --- | --- |
| **Adult Dataset** | **Audit Dataset** | **Immunotherapy Dataset** | **Forest Fires Dataset** |
| **Acc:** 0.86 | **Acc:** 1.0 | **Acc:** 0.81 | **MSE:** 8743.27 |

* 1. **Decision Tree With Distillation:**

1. **Binned Probability Output:**

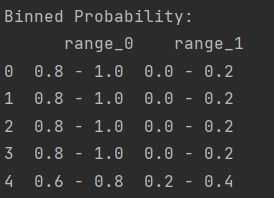
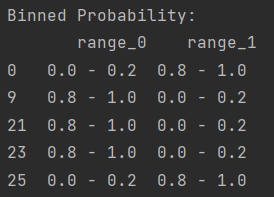
**** **Figure 5 Adult Dataset**

Figure 6 Audit Dataset

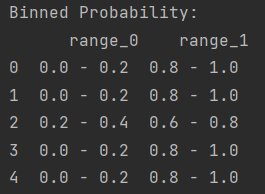


Figure 7 Immunotherapy Dataset

1. **Decision Tree Classifier/Regression:** The output of the model for the given four dataset is shown below in the table.

|  |  |  |  |
| --- | --- | --- | --- |
| **Adult Dataset** | **Audit Dataset** | **Immunotherapy Dataset** | **Forest Fires Dataset** |
| **Acc:** 0.78 | **Acc:** 0.98 | **Acc:** 0.69 | **MSE:** 161.28 |

* 1. **Decision Tree with Original Dataset:** The outputs are shown in the given table.

|  |  |  |  |
| --- | --- | --- | --- |
| **Adult Dataset** | **Audit Dataset** | **Immunotherapy Dataset** | **Forest Fires Dataset** |
| **Acc:** 0.85 | **Acc:** 1.0 | **Acc:** 0.81 | **MSE:** 10204 |

* 1. **Decision Tree Graph:** Including few of the decision tree graphs for some of the dataset.

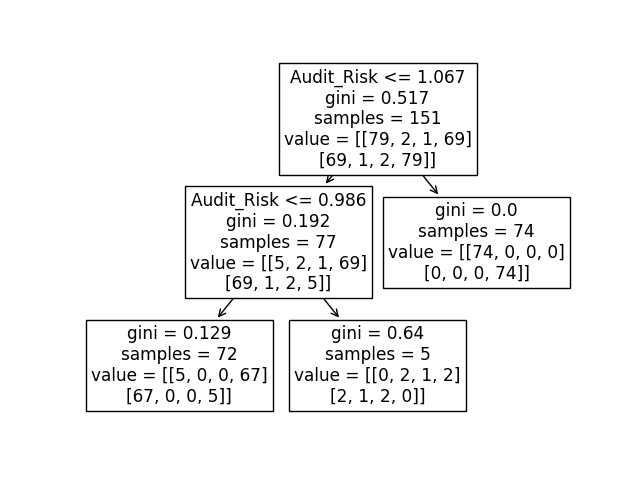
****

Figure 8 Audit Distillation Decision Tree Graph

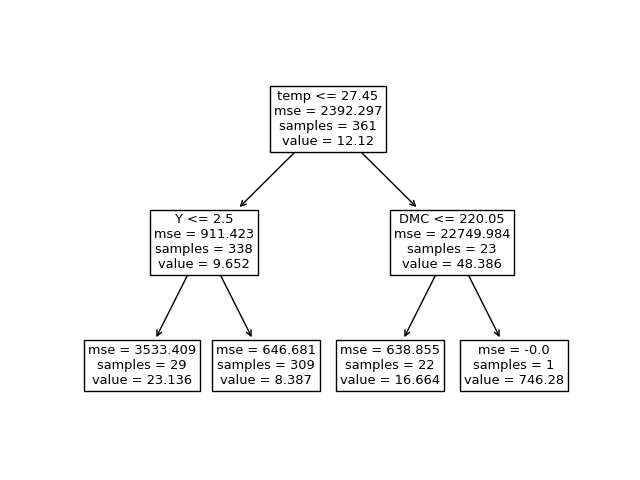
****

Figure 9 Forest Fire Decision Tree Graph

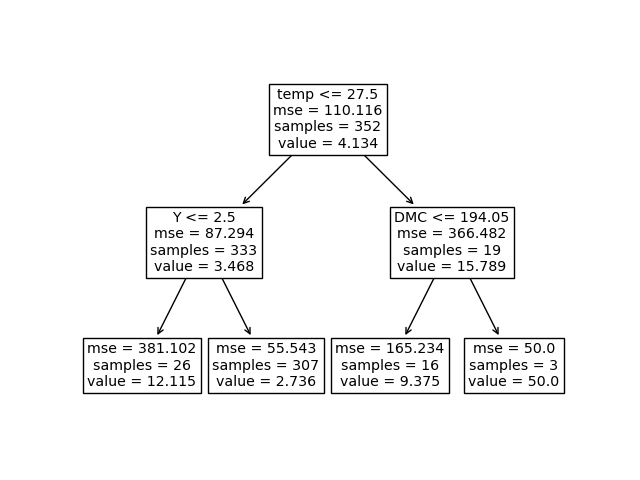


Figure 10 Forest Fire Distillation Decision Tree Graph

1. **DISCUSSION**

From the above results it is clearly visible that the binary classification model performs less accurate using distillation process then the classification model with the original dataset. But for the continuous target variable or for regression decision tree model it performs way better as it reduce the continuous variable range to bins and thus act more accurate. Even, the computation cost is reduced when working with continuous variable.

If we have used distillation knowledge in feature it would have improved the accuracy of binary classification model, as the feature values varies a lot and noise and outliers could be detected and removed using the same knowledge.

However, there many challenges when considering distillation knowledge to be used in machine learning models: distillation type and knowledge quality. Here, we deduce a model smoothing and regularization using label based distillation type, but the type of distillation quality faces a major challenge in the given task.

1. **CONCLUSION**

Distillation process can improve a model computational cost and accuracy by binding the target values into specific range. But, it doesn’t perform well for the binary classification model as the target value consists of 2 unique values. From our task, it is shown that the random forest classifier with 500 trees, have performed better than the decision tree classifier with and without distillation process. However, for continuous target variable decision tree regressor with distillation knowledge has performed a lot better than the random forest classifier and decision tree model without distillation knowledge. In future, we would try to improve binary classification model accuracy using distillation knowledge using feature knowledge and even try to improve to quality of knowledge so get a perfect distillation decision tree model.

## REFERENCES

[1] “UCI Machine Learning Repository,” *Uci.edu*, 2018. <https://archive.ics.uci.edu/ml/index.php>.

[2] “UCI Machine Learning Repository: Adult Data Set,” *Uci.edu*, 2019. <https://archive.ics.uci.edu/ml/datasets/Adult>.

[3] “UCI Machine Learning Repository: Audit Data Data Set,” *archive.ics.uci.edu*. <https://archive.ics.uci.edu/ml/datasets/Audit+Data>.

[4] “UCI Machine Learning Repository: Immunotherapy Dataset Data Set,” *archive.ics.uci.edu*. <https://archive.ics.uci.edu/ml/datasets/Immunotherapy+Dataset>.

[5] “UCI Machine Learning Repository: Forest Fires Data Set,” *archive.ics.uci.edu*. <https://archive.ics.uci.edu/ml/datasets/Forest+Fires>.

[6] J. Gou, B. Yu, S. Maybank, and D. Tao, “Knowledge Distillation: A Survey,” *Int. J. Comput. Vis.*, 2021, doi: 10.1007/s11263-021-01453-z.

[7] J. Gou and B. Yu, “Knowledge Distillation: A Survey | Request PDF,” *ResearchGate*, Mar. 03, 2021. <https://www.researchgate.net/publication/350293589_Knowledge_Distillation_A_Survey>.

[8] F. Khozeimeh, R. Alizadehsani, M. Roshanzamir, A. Khosravi, P. Layegh, and S. Nahavandi, "An expert system for selecting wart treatment method," Computers in Biology and Medicine, vol. 81, pp. 167-175, 2/1/ 2017.

[9] F. Khozeimeh, F. Jabbari Azad, Y. MahboubiOskouei, M. Jafari, S. Tehranian, R. Alizadehsani, et al., "Intralesional immunotherapy compared to cryotherapy in the treatment of warts," International Journal of Dermatology, 2017, DOI: 10.1111/ijd.13535