Automated Machine Learning

Manikantha K[1], Shreyans Jain[2], Shreyas K Shankar[3], Sachin R Bujari[4], Suraj C T[5]

[1]Assistant Professor, Dept of CSE, BNMIT, Karnataka, India
[2] UG student, Dept of CSE, BNMIT, Karnataka, India
3 UG student, Dept of CSE, BNMIT, Karnataka, India
4 UG student, Dept of CSE, BNMIT, Karnataka, India
5UG student, Dept of CSE, BNMIT, Karnataka, India

ABSTRACT

Automated machine learning, also referred to as AutoML is the process of automating the time consuming, iterative tasks of machine learning model development. Here the user need not have any knowledge of machine learning or python. In this paper we propose a web-based AutoML system which considerably reduces user intervention. The user has to just upload the dataset and mention the target column, which is taken by our system as the input and a model file is generated as the output which can be used by the user for prediction. In this system we use three different techniques namely Regression, Classification and Neural Networks for generating the best machine learning model for the uploaded dataset.

Keyword: - *Data cleaning, Feature engineering, Hyper-parameter tuning, Lasso Regression, Ridge Regression, Logistic Regression, Neural Network*

1. INTRODUCTION

Automated machine learning (AutoML) is the process of automating the end-to-end process of applying machine learning to real-world problems. AutoML makes machine learning available in a true sense, even to people with no major expertise in this field. It allows data scientists, analysts and developers to build ML models with high scale, efficiency and productivity all while sustaining model quality. The service then iterates through ML algorithms paired with feature selections, where each iteration produces a model with a training score. The higher the score, the better the model is considered to "fit" the data. AutoML makes machine learning more accessible by automatically generating a data analysis pipeline that can include data pre-processing, feature selection and feature engineering methods along with machine learning methods and parameter settings that are optimized for a given data.
The main steps involved in the proposed system are:

Reading the input data: The input data is retrieved from the source and checked for any redundant or null values.

Pre-processing: Pre-processing the data includes feature extraction, feature selection, feature engineering and feature optimization.

Optimization: Optimization involves selecting the correct model for our dataset in order to produce the required output. This involves algorithm selection and hyperparameter optimization.

Evaluation: The obtained model should be evaluated with the user data set which results in higher model accuracy and efficient parameter tuning.

1.1 SYSTEM DESIGN

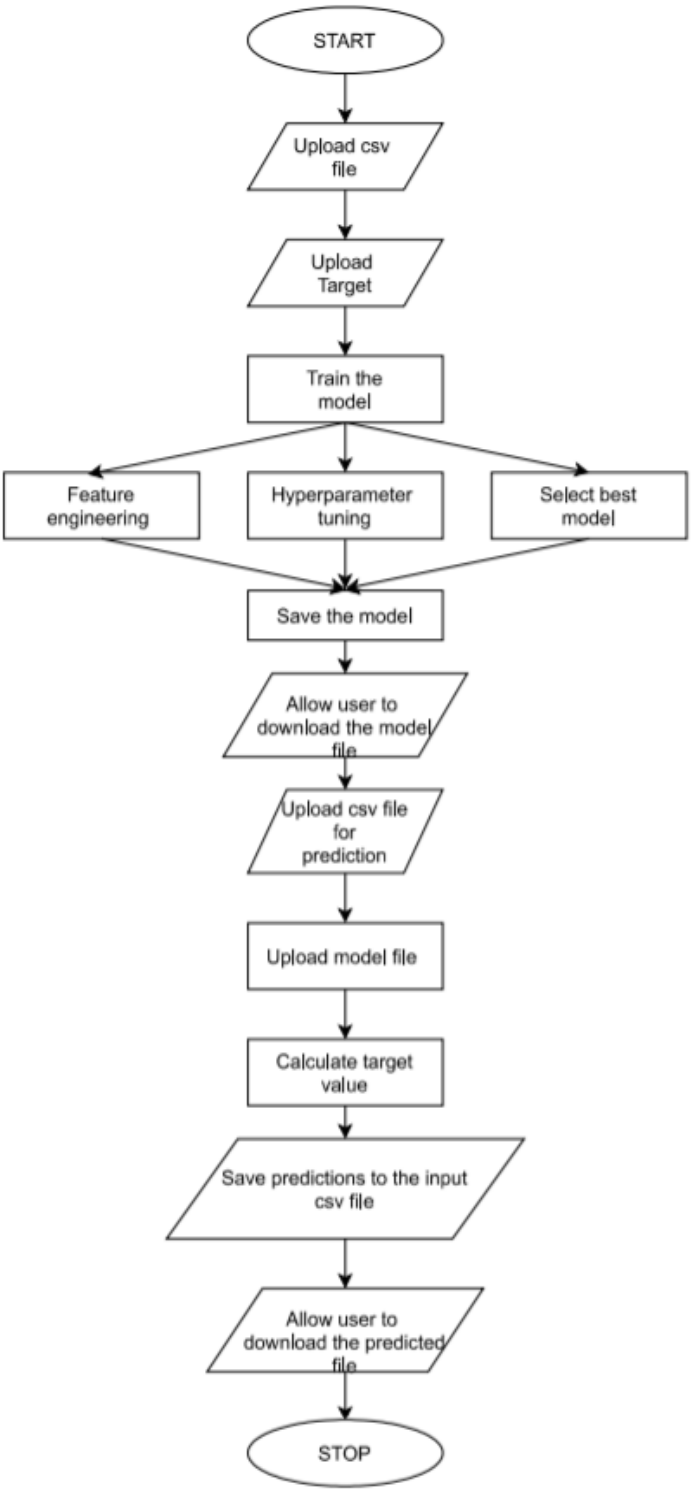

Figure 1: System Architecture

The main goal of the proposed system is to develop an AutoML pipeline in python which allows the user to enter a dataset and provides the best model which fits that particular dataset using machine learning techniques. The obtained model can be used to predict the values for any new instances of the trained dataset.

1.2 Data Pre-processing:

Data Pre-processing is that step in which the data gets transformed, or Encoded, to bring it to such a state that now the machine can easily parse it. In other words, the features of the data can now be easily interpreted by the algorithm.

1.3 Feature Engineering:

Feature Engineering is the step which deals with the preparing the proper input dataset, compatible with the machine learning algorithm requirements and improving the performance of machine learning models. Basically it will have the features from the data which are not relevant or contributing very much to predictions and which helps to reduce the hypothesis space and make our model's job easier to learn faster.

1.4 Hyper parameter Tuning:

Hyper Parameter Tuning is choosing a set of optimal hyper parameters for a learning algorithm. In this method searches for the best hyper parameters for an algorithm using the methods such as Grid Search, Random Search and Bayesian optimization.

Model Selection:

Model selection is the task of selecting a statistical model from a set of candidate models, for the given data. All the algorithms will be trained with same data and the accuracy is calculated for each model. Then the trained model with the best accuracy score will be selected.

Prediction: User can upload the Dataset for which the result has to be predicted, the best model which is selected will be used for computing the result. Result will be given to the user.

2. IMPLEMENTATION

Three techniques are used to implement the proposed system namely, Classification, Regression and Neural Network.

Classification

Classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known.

i.Random Forest Classifier:

Random forests is a supervised learning algorithm. It consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest classifier outputs the class with the most votes, which becomes the model's prediction. The low correlation between models is the key.

ii.Support Vector Machine (SVM):

A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier.

iii.Logistic Regression:

Logistic regression is a statistical method for predicting binary classes. The outcome or target variable is dichotomous in nature. It is a special case of linear regression where the target variable is categorical in nature. It uses a log of odds as the dependent variable. Logistic Regression predicts the probability of occurrence of a binary event utilizing a logit function.

2.1 Artificial Neural Network Classifier

Artificial Neural Networks (ANNs) can be used as a nonparametric technique for classification. They differ significantly from the level-slice and histogram estimation algorithms in that the decision boundaries are not fixed by a deterministic rule applied to the prototype training signatures but are determined in an iterative fashion by minimizing an error criterion on the labelling of the training data. The neural network is given the target outputs on to which it should map its inputs, i.e. it is given in paired data of input and output. The error arising from the discrepancy between the network output and the target is used to optimize the network parameters. Once the network has been trained, it is used to produce an output for unseen data.

Neural Network is implemented using Tensor Flow. It is an open source software library for numerical computation using dataflow graphs. Nodes in the graph represents mathematical operations, while graph edges represent multi-dimensional data arrays (aka tensors) communicated between them. The flexible architecture allows to deploy computation to one or more CPUs or GPUs in a desktop, server or mobile device with a single API.

Regression

Regression takes a group of random variables, thought to be predicting a target variable, and tries to find a mathematical relationship between them. This relationship is typically in the form of a straight line that best approximates all the individual data points.

Linear Regression:

Linear regression is a linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression.

Lasso Regression:

Least Absolute Shrinkage and Selection Operator (Lasso) regression is an example of regularized regression. Regularization is one approach to tackle the problem of over fitting by adding additional information and thereby shrinking the parameter values of the model to induce a penalty against complexity. Lasso regression can result in feature selection. Lasso regression can completely eliminate the variable by reducing its coefficient value to 0.

Ridge Regression:

The Ridge regression is a technique which is specialized to analyze multiple regression data which is multi collinearity in nature. Ridge regression is a method that seeks to reduce the MSE by adding some bias and, at the same time, reducing the variance.

3. Experimental Results

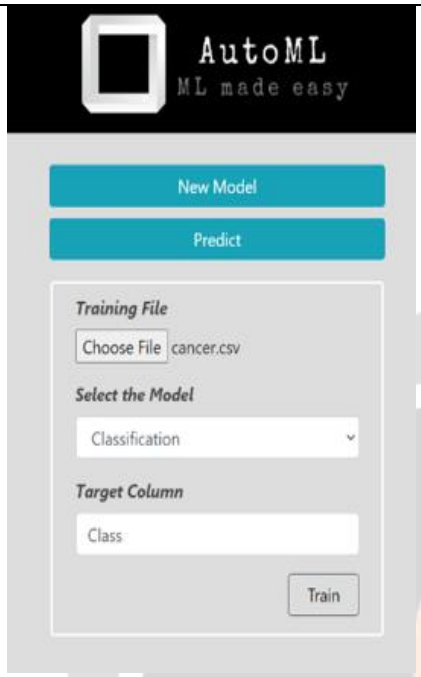

Figure 2: Input dataset and target column to the system

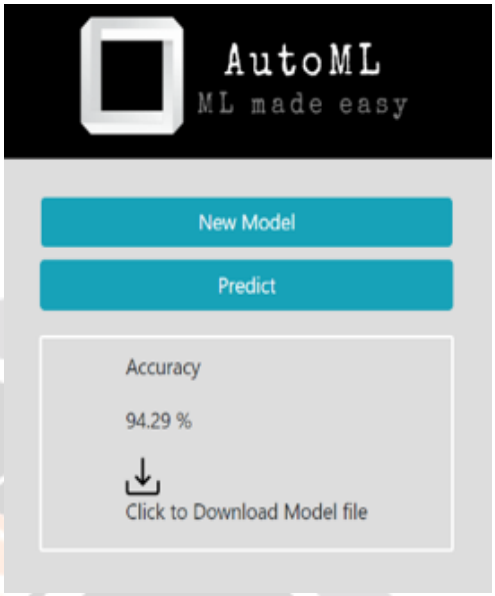

Figure 3: Model generated by the system

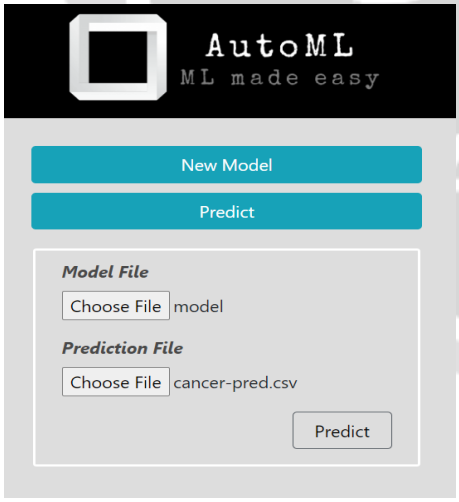

Figure 4 : Input model and prediction file to the system

Figure 5: Predicted file generated by the system

4. PERFORMANCE EVALUATION

Classification

Confusion matrix:

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.

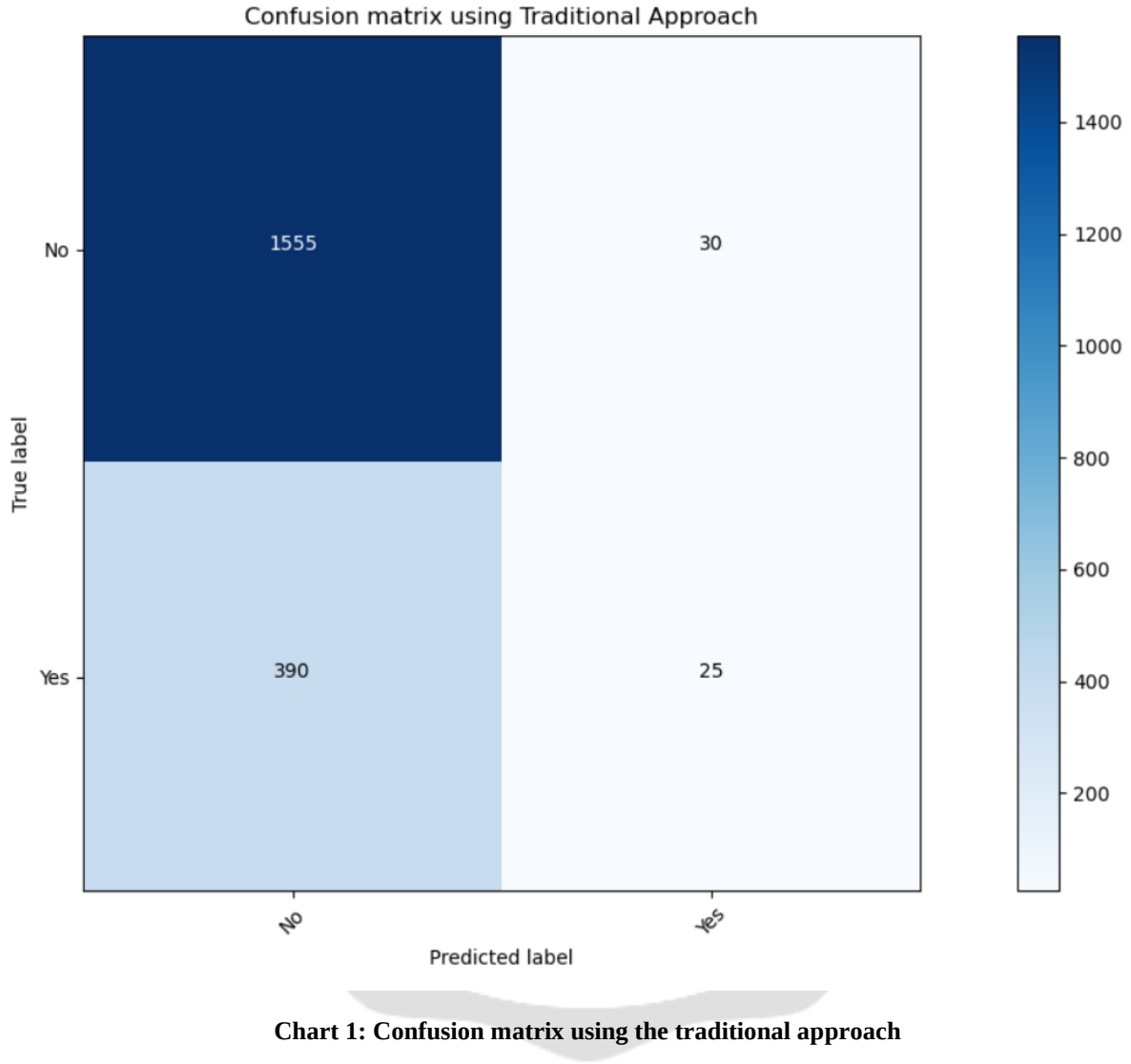

Chart 1: Confusion matrix using the traditional approach

The number in each cell indicates the number of instances in the test data with the corresponding predicted classes. The number of true positives and true negatives gives an idea of how well classifier is performing. In the above matrix true negatives are 1555 and true positives are 25, so the accuracy of this classifier is 79%.

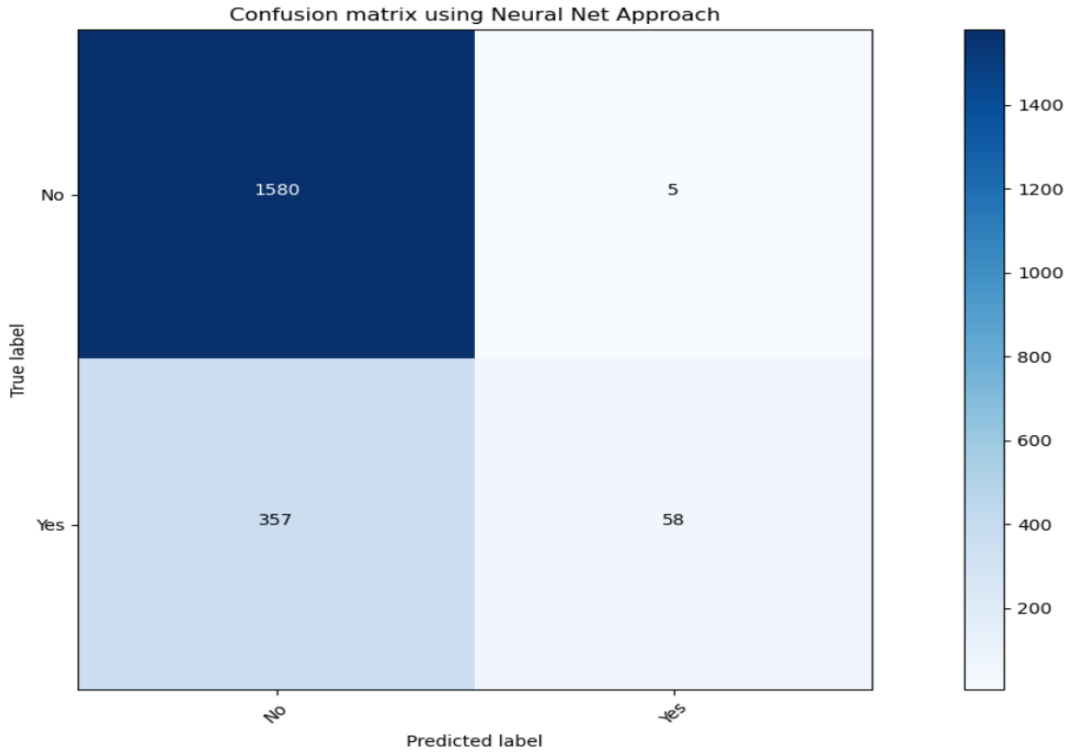

Chart 2: Confusion matrix using neural network

In the above matrix true negatives are 1580 and true positives are 58, so the accuracy of this classifier is 81.9%. The time taken for the classifier to generate the result is low and also the number of examples which are misclassified have decreased. So the neural network approach shows the better performance. It is advisable to use neural network approach over the traditional approach to save time while training and also to obtain better model.

Regression

The performance of a regression model can be understood by knowing the error rate of the predictions made by the model. Performance can also be measured the by knowing how well the regression line fit the dataset.

Below is a snapshot of actual values and predicted values in a weather forecast dataset. Here the column labelled "FR_windspeed_10m" is the target column.

utc_timestamp	FR_temperature	FR_radiation_direct_horizontal	FR_radiation_diffuse_horizontal	FR_windspeed_10m
2010-01-01T06:00:00Z	1.232	3.83E-06	0.000296172	5.42
2010-01-01T07:00:00Z	1.055	0.033623198	1.879676802	5.43
2010-01-01T08:00:00Z	1.212	1.240614993	31.28378501	5.55
2010-01-01T09:00:00Z	1.725	6.808989721	84.41671028	5.89
2010-01-01T10:00:00Z	2.254	15.44580797	126.871792	6.1

Table 1: Dataset with actual target values

utc_timestamp	FR_temperature	FR_radiation_direct_horizontal	FR_radiation_diffuse_horizontal	FR_windspeed_10m
2010-01-01T06:00:00Z	1.232	3.83E-06	0.000296172	4.787897043
2010-01-01T07:00:00Z	1.055	0.033623198	1.879676802	4.730349394
2010-01-01T08:00:00Z	1.212	1.240614993	31.28378501	4.985435065
2010-01-01T09:00:00Z	1.725	6.808989721	84.41671028	5.495514431
2010-01-01T10:00:00Z	2.254	15.44580797	126.871792	5.89845792

Table 2: Dataset with predicted target values

To measure performance of the regression model based on the above datasets, two of the most popular metrics are discussed. They are-

- Mean Absolute Error(MAE)
- Root Mean Square Error (RMSE)

Mean Absolute Error(MAE)

This is the simplest of all the metrics. It is measured by taking the average of the absolute difference between actual values and the predictions.

$$MAE=\frac{1}{n}\sum|y-\hat{y}|$$

If we take the absolute difference of the predictions and the actual values we get:

|5.42-4.78|=0.64

|5.44-4.73|=0.71

|5.55-4.98|=0.57

|5.89-5.49|=0.4

Root Mean Square Error (RMSE)

The Root Mean Square Error is measured by taking the square root of the average of the squared difference between the prediction and the actual value. It represents the sample standard deviation of the differences between predicted values and observed values (also called residuals). It is calculated using the following formula:

$$RMSE=\sqrt{\frac{\sum_{i=1}^{N}(Predicted_i-Actual_i)^2}{N}}$$

If we take our previous example again, then we have:

RMSE = sqrt ((0.41+0.5+0.32+0.16+0.04) / 5) = 0.53

Here the RMSE value is greater than that of MAE. This is because RMSE takes the square of the differences between the predictions and the actual value, hence the value is greater the MAE value.

5. CONCLUSIONS

Machine learning and AI are becoming more accessible with each passing year. While high-level libraries like Keras hide the underlying complexity of deep learning models, AutoML approaches take one step further and are able to provide feasible machine learning models just from a dataset as an input. This provides a smooth pathway into machine learning for non-experts. AutoML can provide production-ready models for small start-ups that cannot dedicate enough budget to hiring ML experts.

ThisdoesnotmeanthatAutoMLisonlydirectedtowardsnon-experts.Techniquesused in AutoML libraries can provide powerful tools for automated optimization for developers and the results of AutoML searches can provide valuable intuition towards model choices and hyper parameter configurations. AutoML also does not mean that there'll be no need for machine learning experts – collecting data, ingesting data, cleaning and pre-processing, monitoring and evaluating are important parts of any ML pipeline and require expertise. At the end, towards the aim of making AI more available to the general public, developments in AutoML constitute a huge stride in the right direction. With the recentrise it's seen as a research interest, AutoML can revolutionize the way ML is practiced.

6. REFERENCES

[1] Xin He, Kaiyong Zhao, Xiaowen Chu on topic "AutoML: A Survey of the State-of-the-Art" by Department of Computer Science, Hong Kong Baptist University, August-2019.
[2] Mohamed Maher, Sherif Sakr on topic "SmartML: A Meta Learning-Based Framework for Automated Selection and Hyperparameter Tuning for Machine Learning Algorithms" Published in Proceedings of the 22nd International Conference on Extending Database Technology (EDBT), March 26-29, 2019.
[3] Quanming Yao, Mengshuo Wang, Yuqiang Chen, Wenyuan Dai, Yi-Qi Hu, Yu-Feng Li, Wei-Wei Tu, Qiang Yang, Yang Yu on topic "Taking the Human out of Learning Applications: A Survey on Automated Machine Learning" Published in arxiv.org, January-2019.
[4] Janek Thomas, Stefan Coors, Bernd Bischl on topic "Automatic Gradient Boosting" published by Department of Statistics, LMU, Ludwigstrasse 33, D80539 Munich ,13 July 2018.
[5] Peter Flach on topic "Performance Evaluation in Machine Learning: The Good, the Bad, the Ugly, and the Way Forward" published by The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19), July 2019.