# OPTIMIZE EXTERNAL ASSISTANCE FOR ML MODEL USING AUTOML CONCEPT

**Dr. Javed Akhtar Khan**

Department of Computer Science & Engineering
Indur Institute of Engineering & Technology

***Abstract :-*** *As we know in this era machine learning is widely use for various application to train the and test the data set . Now a day many model are available for the particular application or its data set .Many libraries are available to optimize our task in training point of view and testing point of view it means machine learning models has never been major critical to use and access .Here one major challenging task is select the model , train the model as per available data set , perform various operation in the conventional mode, these all the thing very time consuming and sometime create the error also . So avoid this new research concerned with auto fitting machine learning model is introduce . The main concept behind this to optimize the External Assistance for train the ML model for particular application. So here this paper cover the case study of AutoML with its impact over the application .*

***Key words:-*****Machine Learning , ML model , Training Data set , Testing Data Set .**

## 1-Introduction

Data analysis is a powerful tool for learning insights on how to improve the decision making, business model and even products. This involves the construction and training of a machine learning model which faces several challenges due to lack of expert knowledge. This challenges can be over come by using automated machine learning(AutoML) field. AutoML refers to the process of studying a traditional machine learning model development pipeline to segment it into modules and automate each of those to accelerate workflow. With the advent of deeper models, such as the ones used in image processing, Natural Language Processing, etc., there is an increasing need for tailored models that can be crafted for specific workloads. However, such specific models require immense resources such as high capacity memory, strong GPUs, domain experts to help during the development and long wait times during training. The task gets critical as there is not much work done for creating a formal framework for deciding model parameters without the need for trial and error. These nuances emphasized the need for AutoML where automation can reduce turnaround times and also increase the accuracy of the derived models by removing human errors. In recent years, several tools and models have been proposed in the domain of AutoML. Some of these focus on particular segments of AutoML such as feature engineering or model selection, whereas some models attempt to optimize the complete pipeline. These tools have matured enough to be able to compare with human experts on Kaggle competitions and at times have beat them as well, showcasing their veracity.

## 2.AUTOML

AutoML such as autonomic cloud computing, Intelligent Vehicular networks, Block Chain,Software Defined Networking, among others. This paper aims at providing an overview of the advances seen in the realm of AutoML in recent years. We focus on individual aspects of AutoML and summarize the improvements achieved in recent years. The motivation of this paper stems from the unavailability of a compact study of the current state of AutoML. While we acknowledge the existence of other surveys, their motive is to either provide an in-depth understanding of a particular segment of AutoML, provide just an experimental comparison of various tools used or are fixated towards deep learning models. There

is a lot of buzz for machine learning algorithms as well as a requirement for its experts. We all know that there is a significant gap in the skill requirement.

## 3.H2O ARCHITECTURE

The motive of H2O is to provide a platform which made easy for the non-experts to do experiments with machine learning. H2O architecture can be divided into different layers in which the top layer will be different APIs, and the bottom layer will be H2O JVM. H2O's core code is written in Java that enables the whole framework for multi-threading. Although it is written in Java, it provides interfaces for R, Python and few others shown in the architecture, thus enabling it to be used efficiently. In crux, we can say that H2O is an open source, in memory, distributed, fast and scalable machine learning and predictive analytics that allow building machine learning models to be an ease.If you are using python the same method is applied in it too, from segment of AutoML, provide just an experimental comparison of various tools used or are fixated towards deep learning models There is a lot of buzz for machine learning algorithms as well as a requirement for its experts. We all know that there is a significant gap in the skill requirement. The motive of H2O is to provide a platform which made easy for the non-experts to do experiments with machine learning. H2O architecture can be divided into different layers in which the top layer will be different APIs, and the bottom layer will be H2O JVM. H2O's core code is written in Java that enables the whole framework for multi-threading. Although it is written in Java, it provides interfaces for R, Python and few others shown in the architecture, thus enabling it to be used efficiently. In crux, we can say that H2O is an open source, in memory, distributed, fast and scalable machine learning and predictive analytics that allow building machine learning models to be an ease.
Now talking about AutoML part of H2O, AutoML helps in automatic training and tuning of many models within a user-specified time limit. The current version of AutoML function can train and cross-validate Random Forest, an Extremely-Randomized Forest, a random grid of Gradient Boosting Machines (GBMs), a random grid of Deep Neural Nets, and then trains a Stacked Ensemble using all of the models. When we say AutoML, it should cater to the aspects of data preparation, Model generation, and Ensembles and also provide few parameters as possible so that users can perform tasks with much less confusion. H2o AutoML does perform this task with ease and the minimal parameter passed by the user.In both R and Python API, it uses the same data related arguments x, y, training_frame, validation frame out of which y and training_frame are required parameter and rest are optional. You can also configure values for max_runtime_sec and max_models here max_runtime_sec parameter is required, and max_model is optional if you don't pass any parameter it takes NULL by default. The x parameter is the vector of predictors from training_frame if you don't want to use all predictors from the frame you passed you can set it by passing it to x.

## 4. PARAMETERS

Now let's talk about some optional and miscellaneous parameters, try to tweak the parameters even if you don't know about it, it will lead you to gain knowledge over some advanced topics:
Validation frame: This parameter is used for early stopping of individual models in the automl. It is a data frame that you pass for validation of a model or can be a part of training data if not passed by you.
Leaderboard_frame: If passed the models will be scored according to the values instead of using cross-validation metrics. Again the values are a part of training data if not passed by you.
nfolds: K-fold cross-validation by default , can be used to decrease the model performance.
Fold_columns: Specifies the index for cross-validation.
Weights_column: If you want to provide weights to specific columns you can use this parameter, assigning weight 0 means you are excluding the column.
Ignored_columns: Only in python, it is converse of x.
Stopping_metric: Specifies a metric for early stopping of the grid searches and models default value is logloss for classification and deviation for regression.
Sort_metric: The parameter to sort the leaderboard models at the end. This defaults to AUC for binary classification, mean_per_class_error for multinomial classification, and deviance for

regression. The validation_frame and leaderboard_frame depend on the cross-validation parameter that is nfolds.

## 5.Present Model

In the existing model the data preprocess has dine with structured data. Even though data pre-processing consumes a large chunk of time in an ML pipeline, it is astonishing to see the inadequate amount of work done to automate it. For data preprocessing, it can be noted that while the data pre process approaches are adequate for structured data, work still needs to be done to assimilate on Structured data. We suggest the incorporation of data-mining methods as they can deal with such unformed data. This can allow AutoML pipelines to create models capable of learning from Internet sources. In feature engineering, it should be noted that most methods used until now adhere to supervised learning. However, dataset specificity is high, and therefore, AutoML pipelines should be as generic as possible to accommodate the diverse datasets. Therefore, a gradual paradigm shift towards unsupervised.

- Feature Generation is not up to the mark where domain experts excepted results.
- Most AutoML tools emphasize the performance but in the real world, that's just one aspect being covered in machine learning projects. So the companies can't compromise the computing plus storage specification sheet.
- CASH(Combined Algorithm Selection and Hyperparameter) problem considers model selection and hyperparameters optimization as a single hierarchical parameter

## 6. PROPOSED SYSTEM

By using Machine Learning Algorithms we can easily identify a person is fraud or not and get best accuracy. If we are identifying a person is fraud then immediately deactivate a card.

- We segment the AutoML pipeline into parts and review the contributions in each of these segments.
- We explore the various state-of-the-art tools currently available for AutoML and evaluate them.
- We also incorporate the advancements seen in machine learning which seems to be overshadowed by deep learning in recent years.

**Algorithm**:H2O-AutoML, Linear Regression, Gradient Boosting Repressor

## 7.ARCHITECTURE:

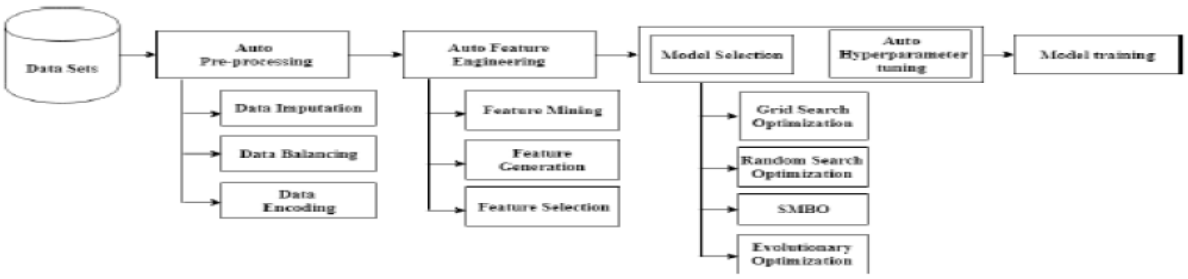


**Figure 1. System Architecture**

## 8.DATA FLOW DIAGRAM:

1. The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of input data to the system, various processing carried out on this data, and the output data is generated by this system.
2. The data flow diagram (DFD) is one of the most important modeling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system.
3. DFD shows how the information moves through the system and how it is modified by a series of transformations.
4. DFD is also known as bubble chart. A DFD may be used to represent a system at any level of abstraction.

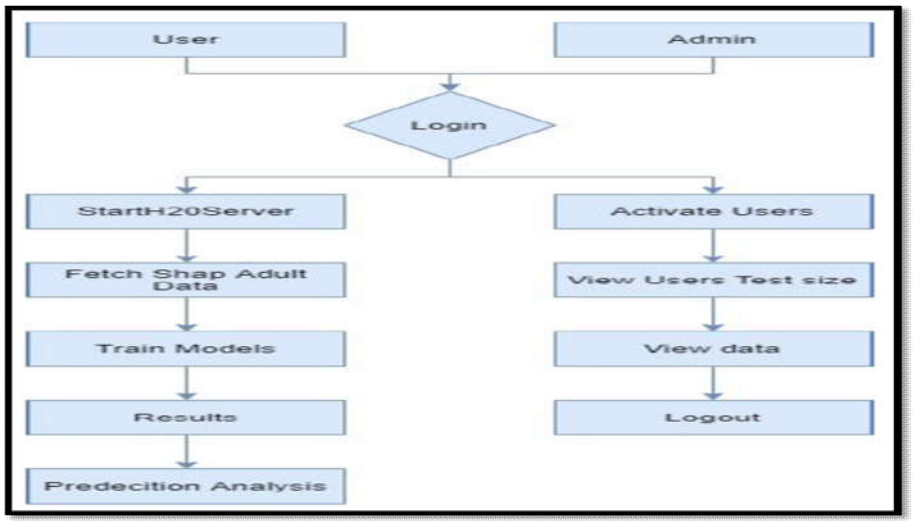


**Figure 2. Data Flow Diagram**

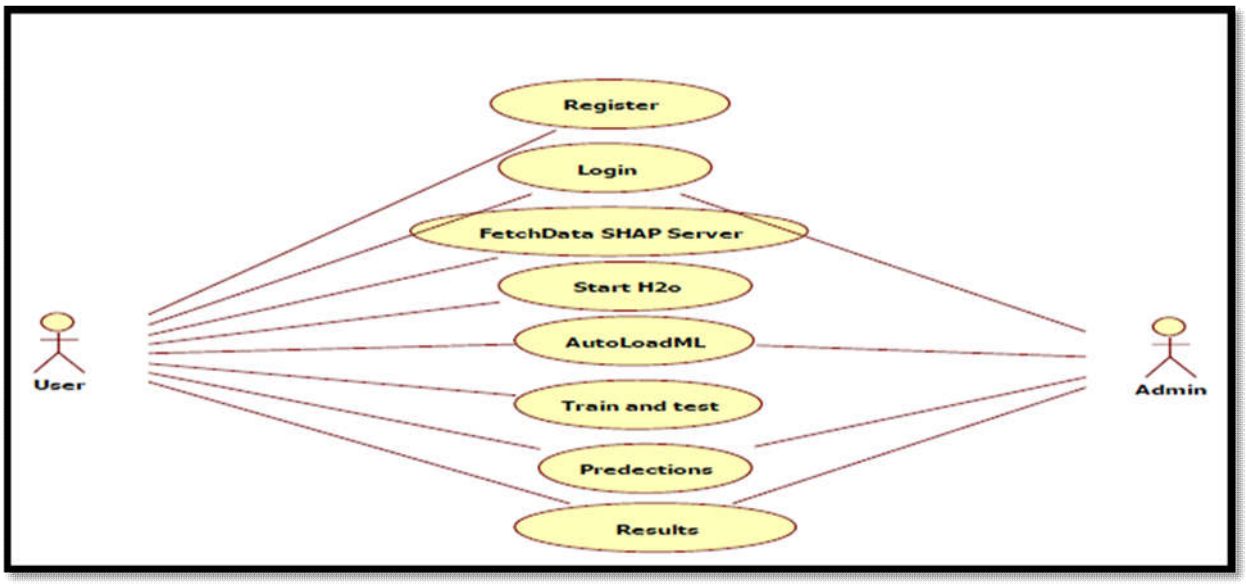


**Figure 3. Use Case Diagram**

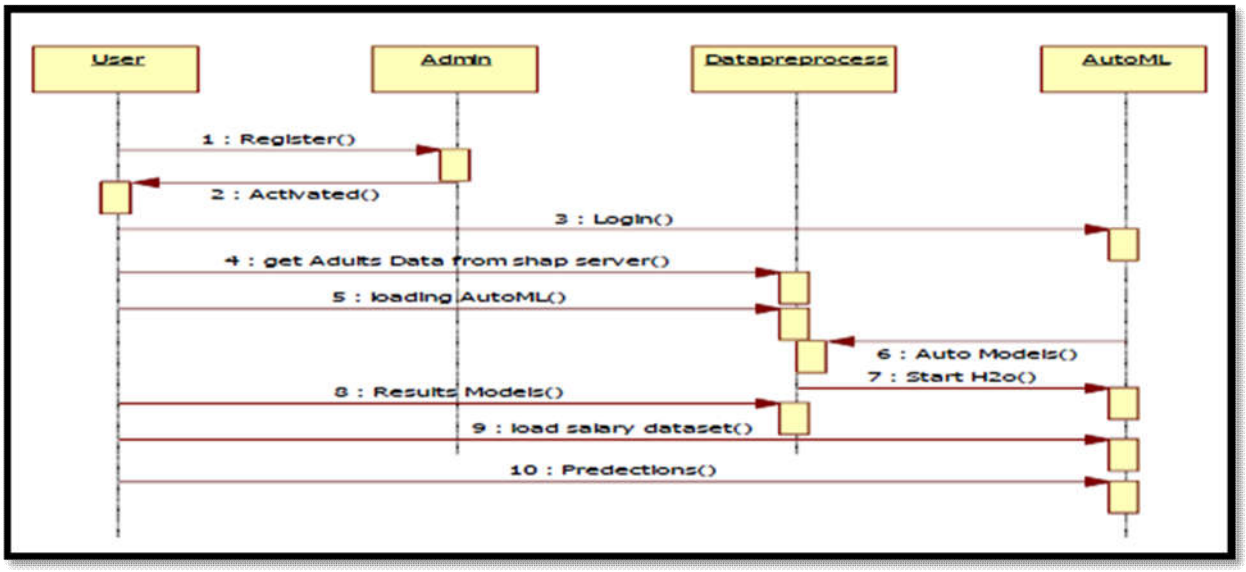


**Figure 4. Sequence Diagram**

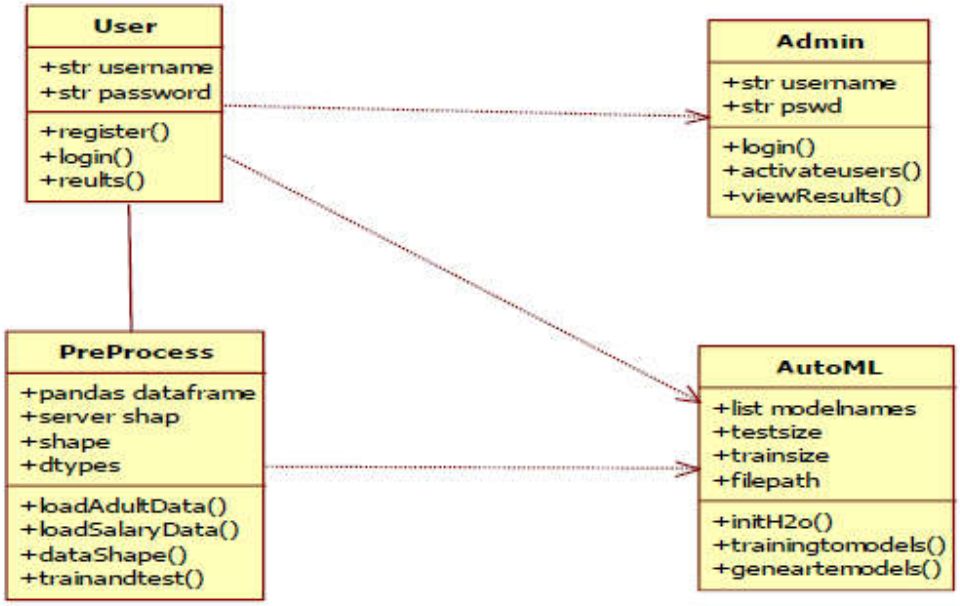


**Figure 5. Class Diagram**

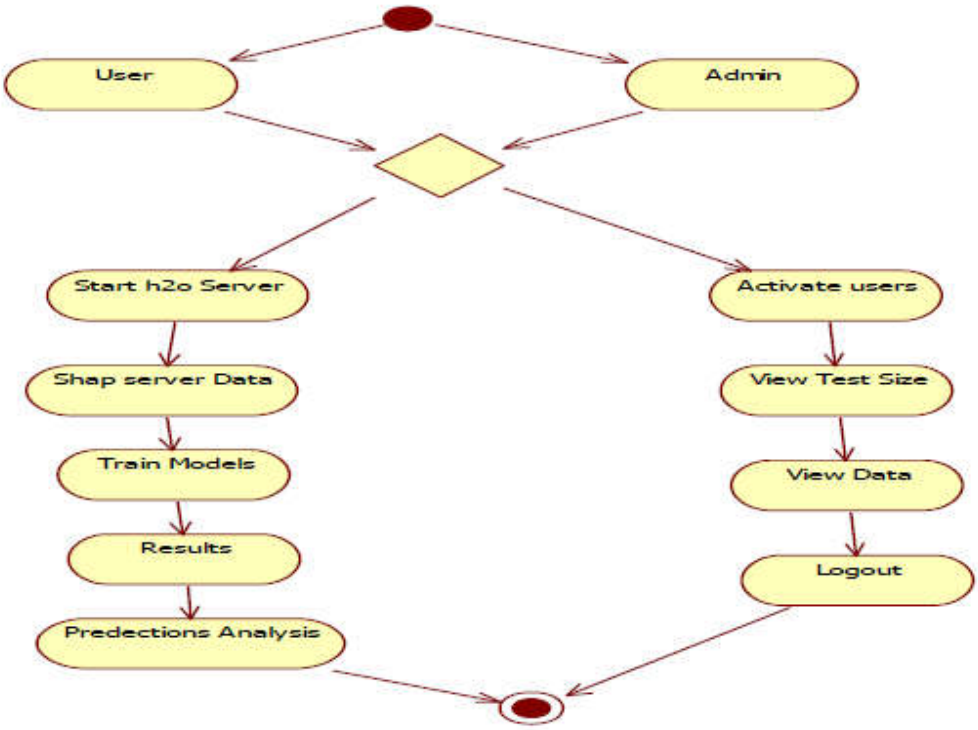


**Figure 6. Activity Diagram**

## 9.RESULT SCREEN

Auto Ml Results

| | model_id | auc | logloss | aucpr | mean_per_class_error | rmse | mse |
|---|---|---|---|---|---|---|---|
| 0 | StackedEnsemble_AllModels_AutoML_20200727_162032 | 0.918305 | 0.308159 | 0.804799 | 0.177644 | 0.309214 | 0.095614 |
| 1 | StackedEnsemble_BestOfFamily_AutoML_20200727_1... | 0.916268 | 0.312327 | 0.801839 | 0.176929 | 0.311859 | 0.097256 |
| 2 | GBM_3_AutoML_20200727_162032 | 0.912290 | 0.413188 | 0.791602 | 0.186498 | 0.357587 | 0.127869 |
| 3 | GBM_4_AutoML_20200727_162032 | 0.911895 | 0.426079 | 0.786651 | 0.178021 | 0.364889 | 0.133144 |
| 4 | GBM_grid__1_AutoML_20200727_162032_model_1 | 0.910124 | 0.346165 | 0.788076 | 0.177121 | 0.323605 | 0.104720 |
| 5 | GBM_1_AutoML_20200727_162032 | 0.909217 | 0.388501 | 0.781646 | 0.191528 | 0.345138 | 0.119121 |
| 6 | GBM_2_AutoML_20200727_162032 | 0.907870 | 0.405938 | 0.767621 | 0.183145 | 0.355160 | 0.126139 |
| 7 | DeepLearning_grid__1_AutoML_20200727_162032_mo... | 0.907017 | 0.318388 | 0.770385 | 0.185953 | 0.319424 | 0.102031 |
| 8 | GLM_1_AutoML_20200727_162032 | 0.904723 | 0.324298 | 0.762277 | 0.189650 | 0.321746 | 0.103521 |
| 9 | DeepLearning_1_AutoML_20200727_162032 | 0.899646 | 0.334766 | 0.753148 | 0.206121 | 0.329101 | 0.108307 |
| 10 | GBM_5_AutoML_20200727_162032 | 0.896027 | 0.463310 | 0.726895 | 0.185710 | 0.385155 | 0.148344 |
| 11 | DRF_1_AutoML_20200727_162032 | 0.872614 | 1.202999 | 0.718824 | 0.207485 | 0.337948 | 0.114209 |
| 12 | XRT_1_AutoML_20200727_162032 | 0.870409 | 0.994401 | 0.719207 | 0.197025 | 0.338989 | 0.114914 |

**Figure 7. Auto saved Server Data**

| S.No | Age | Workclass | EducationNum | MaritalStatus | Occupation | Relationship | Race | Sex | CapitalGain | CapitalLoss | Hoursperweek | Country |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 39.0 | State-gov | 13.0 | Never-married | Adm-clerical | Not-in-family | White | Male | 2174.0 | 0.0 | 40.0 | United-States |
| 2 | 50.0 | Self-emp-not-inc | 13.0 | Married-civ-spouse | Exec-managerial | Husband | White | Male | 0.0 | 0.0 | 13.0 | United-States |
| 3 | 38.0 | Private | 9.0 | Divorced | Handlers-cleaners | Not-in-family | White | Male | 0.0 | 0.0 | 40.0 | United-States |
| 4 | 53.0 | Private | 7.0 | Married-civ-spouse | Handlers-cleaners | Husband | Black | Male | 0.0 | 0.0 | 40.0 | United-States |
| 5 | 28.0 | Private | 13.0 | Married-civ-spouse | Prof-specialty | Wife | Black | Female | 0.0 | 0.0 | 40.0 | Cuba |
| 6 | 37.0 | Private | 14.0 | Married-civ-spouse | Exec-managerial | Wife | White | Female | 0.0 | 0.0 | 40.0 | United-States |
| 7 | 49.0 | Private | 5.0 | Married-spouse-absent | Other-service | Not-in-family | Black | Female | 0.0 | 0.0 | 16.0 | Jamaica |
| 8 | 52.0 | Self-emp-not-inc | 9.0 | Married-civ-spouse | Exec-managerial | Husband | White | Male | 0.0 | 0.0 | 45.0 | United-States |
| 9 | 31.0 | Private | 14.0 | Never-married | Prof-specialty | Not-in-family | White | Female | 14084.0 | 0.0 | 50.0 | United-States |

**Figure 8.1. Performance Prediction Data**

## Perform Predections Analysis

*We love Machine Learning*

| S.No | Years of Experience | Salary |
|---|---|---|
| 1 | 1.1 | 39343.0 |
| 2 | 1.3 | 46205.0 |
| 3 | 1.5 | 37731.0 |
| 4 | 2.0 | 43525.0 |
| 5 | 2.2 | 39891.0 |
| 6 | 2.9 | 56642.0 |
| 7 | 3.0 | 60150.0 |
| 8 | 3.2 | 54445.0 |
| 9 | 3.2 | 64445.0 |
| 10 | 3.7 | 57189.0 |
| 11 | 3.9 | 63218.0 |
| 12 | 4.0 | 55794.0 |
| 13 | 4.0 | 56957.0 |
| 14 | 4.1 | 57081.0 |

**Figure 8.2. Performance Prediction Data**


**Figure 8.3. Performance Prediction Data Plot**

## 10.CONCLUSION

This article cover the various aspect of AUTOML with its some segment ,parameter , H2O is part of AUTOML. Also include the approaches with its bit of explanation and its overview . cover the the some trends and implementation parts ..

## REFERENCES

*[1] Lukas Tuggener, Mohammadreza Amirian, Katharina Rombach, Stefan L¨orwald,Anastasia Varlet, Christian Westermann, and Thilo Stadelmann. Automated machinelearning in practice: state of the art and recent results. In 2019 6th Swiss Conferenceon Data Science (SDS), pages 31–36. IEEE, 2019.*

*[2] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks forlarge-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.*

*[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pretrainingof deep bidirectional transformers for language understanding. arXiv preprintarXiv:1810.04805, 2018.*

*[4] Avatar Jaykrushna, Pathik Patel, Harshal Trivedi, and Jitendra Bhatia. Linearregression assisted prediction based load balancer for cloud computing. In 2018 IEEEPunecon, pages 1–3. IEEE.*

*[5] Jitendra Bhatia, Ruchi Mehta, and Madhuri Bhavsar. Variants of software definednetwork (sdn) based load balancing in cloud computing: A quick review. InInternational Conference on Future Internet Technologies and Trends, pages 164–173. Springer, 2017.*

*[6] Ishan Mistry, Sudeep Tanwar, Sudhanshu Tyagi, and Neeraj Kumar. Blockchain for5g-enabled iot for industrial automation: A systematic review, solutions, andchallenges. Mechanical Systems and Signal Processing, 135:106382, 2020.*

*[7] Jitendra Bhatia, Yash Modi, Sudeep Tanwar, and Madhuri Bhavsar. Softwaredefined vehicular networks: A comprehensive review. International Journal ofCommunication Systems, 32(12):e4005, 2019.*

*[8] Jitendra Bhatia, Ridham Dave, Heta Bhayani, Sudeep Tanwar, and Anand Nayyar.Sdn-based real-time urban traffic analysis in vanet environment. ComputerCommunications, 149:162 – 175, 2020.*

*[9] Xin He, Kaiyong Zhao, and Xiaowen Chu. Automl: A survey of the state-of-the-art.arXiv preprint arXiv:1908.00709, 2019.*
*[10] Radwa Elshawi, Mohamed Maher, and Sherif Sakr. Automated machine learning:State-of-the-art and open challenges. arXiv preprint arXiv:1906.02287, 2019.*
*[11] Anh Truong, Austin Walters, Jeremy Goodsitt, Keegan Hines, Bayan Bruss, andReza Farivar. Towards automated machine learning: Evaluation and comparison of automl approaches and tools. A rXiv preprint arXiv:1908.05557, 2019.*
*[12] Shichao Zhang, Chengqi Zhang, and Qiang Yang. Data preparation for datamining. Applied artificial intelligence, 17(5-6):375–381, 2003.*
*[13] Erhard Rahm and Hong Hai Do. Data cleaning: Problems and current approaches.IEEE Data Eng. Bull., 23(4):3–13, 2000.*
*[14] Dipali Shete and Sachin Bojewar. Auto approach for extracting relevant data usingmachine learning. International Journal of Electronics, 6:0, 2019.*
*[15] Carol M Musil, Camille B Warner, Piyanee Klainin Yobas, and Susan L Jones. Acomparison of imputation techniques for handling missing data. Western Journal of Nursing Research, 24(7):815–829, 2002.*