RESEARCH ARTICLE

RSP Science Hub

International Research Journal on Advanced Science Hub
2582-4376
www.rspsciencehub.com
Vol. 06, Issue 05 May

Check for updates

http://dx.doi.org/10.47392/IRJASH.2024.018

Automated Machine Learning Based Crop Recommendation System

D. Madhu sudhan Reddy[1], *Dr N. Usha Rani*[2]
[1]*Research Scholar, Department of CSE, Sri Venkateswara University, Tirupati, India*
[2]*Associate Professor, Department of CSE, Sri Venkateswara University, Tirupati, India*
Emails: *madhu.dagada@gmail.com*[1], *usharani.ur@gmail.com*[2]

Article history
Received: 23 April 2024
Accepted: 02 May 2024
Published: 30 May 2024

Keywords:
AutoGluon; AutoML; Crop recommendation; H2O; Pycaret

Abstract

Agriculture plays a crucial role in supplying food for the population, as well as contributing significantly to the country's Gross Domestic Product (GDP) in India. For farmers to achieve higher yields and profitability, it is essential to select a crop based on soil parameters. To simplify this process, a system for crop recommendation based on Machine Learning models has been developed. As a result, farmers may find it difficult to make informed decisions when using the Machine Learning approach since it is both time-consuming and exhaustive. Automated Machine Learning is being used to simplify and speed up the process. A machine learning algorithm uses an automatic selection of algorithms, features, and hyperparameters to make predictions, which can result in more accurate results. This study examines various Automated Machine Learning frameworks and compares the accuracy scores of different crop recommendation systems. H2O and AutoGluon achieved the highest accuracy score of 92.0%.

1. Introduction

India holds the top position globally in terms of net cropped area, with the United States and China following closely. The country plays a significant role in the global agricultural industry, with 58% of its population relying on agriculture as their primary source of income. India boasts several remarkable achievements in agriculture, including the largest herd of buffaloes, and extensive cultivation of different crops like wheat, rice, and cotton, and is the world's largest producer of milk, pulses, and spices. Artificial Intelligence, particularly through the implementation of Machine Learning (ML), offers effective solutions to the problems faced by farmers in agriculture. Major challenges faced by Indian farmers in agriculture are, farmers need to identify suitable crops that can thrive in their specific land or soil conditions. Next, there is a lack of automation in the crop cycle system. Lastly, farmers often struggle to obtain fair prices for their agricultural commodities. Precision Agriculture (PA), an emerging field in computer science for agriculture, addresses these issues through the application of artificial intelligence (AI). PA involves the development of AI-based tools and data-driven solutions, such as crop recommendations, fertilizer suggestions, and pest/disease detection, to enhance agricultural outcomes. ML algorithms can assist in selecting the most appropriate crop for specific farming land, taking into account soil parameters such as Nitrogen, Potassium, Phosphorus, and others. By making informed decisions regarding crop selection, farmers can achieve higher yields and increased profits.

2. Related Works

A comparative analysis [1] of the approach of Automated Machine Learning (AutoML), Conventional ensemble learning method and K-

Nearest Oracle AutoML model for predicting the dropout rate of students in sub-Saharan African countries. And results are KNORA AutoML system has given 97% accuracy, and 71% of precision better than the results of the Conventional ensemble model with 96% accuracy, and 70% precision. So AutoML models predict the rate is high. The system [2] can estimate crop yield and biomass using hyperspectral images. The estimation results are provided in the form of determination coefficient (R2) and Normalized Root Mean Square Error (NRMSE) metrics. Under various agricultural resource conditions, the implementation flexibility and learning cost can be reduced by utilizing the open-source system AutoML and an R language-based package. An efficient platform and framework [3] based on AutoML approach H2O to the estimation of Soybean and Corn seed protein and oil composition. The Gradient Boosting Machine(GBM) of H2O outperformed other algorithms for combination images given by Unmanned Aviation Vehicle(UAV) based hyperspectral and LiDAR(Light Detection and Ranging). It also investigated the model with crop images taken at different time points to analyse prediction results. AutoML-based system [4] for weed detection by considering two different datasets produced promising performance results with 93.8% and 90.7% F1 scores depending on the dataset. It also enabled a balance between AutoML and manual expert work to increase efficiency in plant protection. The developed model is evaluated with the original and noisy version dataset of early crop weed and plant seedlings. A comparative analysis [5] of time-series data like stock price, business development, weather, and economic status for prediction through traditional Machine Learning models and AutoML frameworks H2O, AutoSklearn, AutoGluon, TPOT, Autokeras, EvalML, TransmogrifAI along with compare parameters and hyperparameters of concerning Machine Learning algorithms involved in AutoML. Classification system [6] for crops and weeds by two different datasets, dataset 1 collected images from the real world by using the agricultural robot, and dataset 2 is open source available PlantVillage2. AutoML has chosen different CNN-based ensemble models with objective function Dual Metrics(DM) has given better results when compared with other models such as the ensemble model with objective function No Miss Weed (NMW) and Categorical Cross Entropy(CCE).

3. Methods

For this study, a dataset [7] was acquired for the combined district of Kurnool, including Nandyal and Kurnool, located in the state of Andhra Pradesh. The dataset was obtained from an official government website. To ensure data quality, it is necessary to preprocess the dataset by removing duplicate and outlier values. Following the preprocessing step, the dataset consists of a total of 67,788 data records, with 12 columns and 5,649 rows. The dataset exclusively contains soil parameter values, including Potassium (K), Phosphorus (P), Nitrogen (N), pH, Electrical Conductivity (EC), Iron (Fe), Zinc (Zn), Sulphur (S), Organic Carbon (OC), Copper (Cu), Manganese (Mn), and Boron (B).

3.1 Conventional Machine Learning

The conventional process of developing Machine Learning models is manual, requiring substantial domain knowledge, consuming resources, and taking considerable time to yield predictive results. The conventional Machine Learning process involves sequential steps, as depicted in Figure 1 starting from Data Collection, Data Exploration, Data Preparation, Feature Engineering, Model Selection, Model Training, Hyperparameter Tuning, and Prediction.

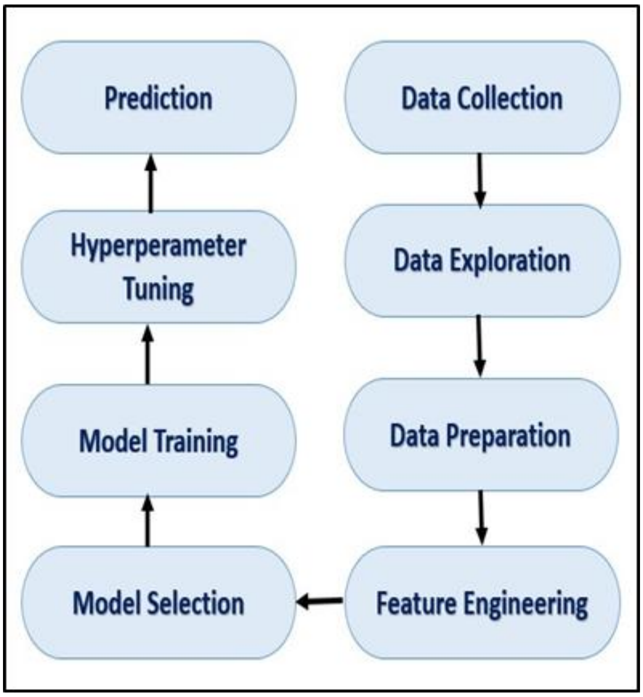

Figure 1 Process of Conventional Machine Learning

The key limitations of conventional Machine Learning are

- It typically requires a high level of expertise in knowledge of feature engineering, algorithm selection, hyperparameter tuning, and model evaluation.
- It often relies on manual feature engineering, where domain experts handcraft features based on their understanding of the problem.
- Conventional Machine Learning requires manual tuning of hyperparameters, which can be challenging and time-consuming due to the large parameter space.
- It relies on a limited set of popular algorithms that are well-known and well-understood by researchers and practitioners. Exploring alternative algorithms requires manual effort and expertise.
- Conventional ML workflows are often not easily reproducible, as the manual nature of the process can lead to inconsistencies.
- Conventional ML frameworks and tools may have a steep learning curve, making it difficult for non-experts to leverage ML techniques effectively.

3.2 Automated Machine Learning

The term "Automated Machine Learning," also known as “AutoML” [8] refers to the utilization of techniques, procedures, and frameworks to automate all or part of the Machine Learning pipeline. It provides pre-built components and resources to expedite and enhance the machine-learning process. AutoML streamlines the model development process by minimizing concerns about specific implementation details like hyperparameters, individual model selection, and other minor aspects that could impede progress. With AutoML, the creation of production-ready Machine Learning models becomes faster, simpler, and more efficient. AutoML aims to establish machine learning by making it accessible to users with limited expertise in data science or programming. It allows individuals from various domains to leverage the power of machine learning without the need to have an in-depth understanding of its intricacies. AutoML tools and platforms often provide user-friendly interfaces and workflows to streamline the machine learning process and enable faster development and deployment of models. The steps involved in Auto-ML are illustrated in Figure 2.

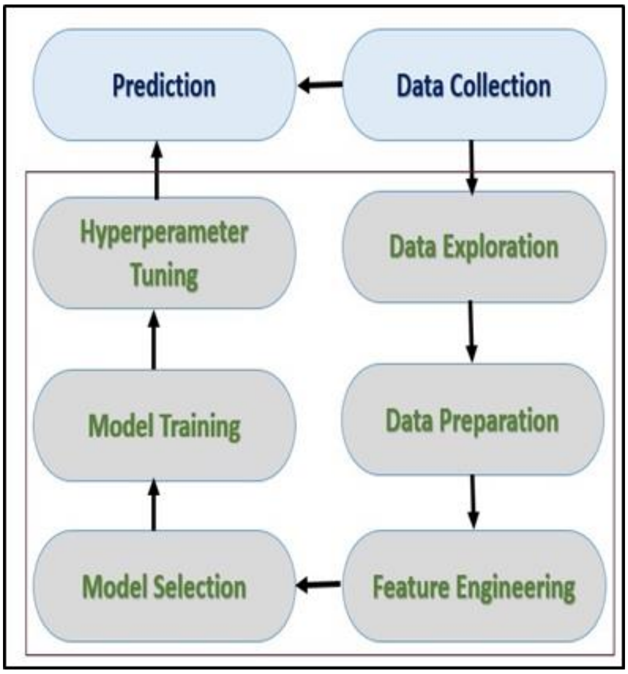

Figure 2 Process of Automated Machine Learning

One of the key aspects of AutoML is data preprocessing. This involves handling various data-related tasks such as missing data imputation, outlier detection, and data normalization or scaling. AutoML algorithms automatically handle these tasks, ensuring that the data is in a suitable format for model training. Another important step is feature engineering. Traditionally, feature engineering requires domain experts to manually identify and create relevant features from raw data. However, AutoML algorithms can automatically select useful features from the available data or generate new features through transformations or combinations. This reduces the need for manual feature engineering and can potentially lead to more efficient and accurate models. Model selection is another crucial aspect that AutoML addresses. Instead of manually selecting and evaluating different models, AutoML tools explore a wide range of models, including both traditional algorithms and more advanced techniques like neural networks. They automatically evaluate the performance of these models on the given dataset and problem, helping users identify the best-performing model without requiring extensive knowledge of various algorithms. Hyperparameter tuning is a critical step in optimizing model performance. AutoML automates this process by systematically searching the hyperparameter space to find the optimal configuration for a given model and dataset. Techniques like random search, Bayesian optimization, grid search, or genetic algorithms are employed to explore the hyperparameter space and find the best combinations efficiently. Furthermore, AutoML provides mechanisms for model evaluation and selection. It allows users to assess the performance of different models using evaluation metrics and validation techniques. This helps users make informed decisions about which model to choose for deployment based on their specific requirements and performance criteria. AutoML can be applied to various domains, including natural language processing, computer vision, and other deep learning frameworks. It automates the challenging tasks of selecting the appropriate model and improving its performance based on the provided data. AutoML simplifies the Machine Learning process, making it more accessible and resembling a black box. It automates various stages of the ML pipeline that involve applying algorithms to real-world situations. Typically, a human operator would require an understanding of the algorithm's internal logic and its practical application. Some of the AutoML frameworks such as AutoKeras, AutoSklearn, AutoGluon, Pycaret, H2O, MLBox, AutoWeka, and TransmogrifAI.

3.2.1 H2O

H2O [9] is open source and created by H2O.ai and it is distributed in-memory Machine Learning platform. R and Python are both supported by H2O. It supports the most popular statistical and Machine Learning methods, such as deep learning, generalised linear models, gradient-boosted machines, etc, H2O employs its algorithms to build pipelines and contains a module for AutoML. Pipelines are optimised using a thorough search for feature engineering techniques and hyperparameter tuning. H2O automates major activities like feature engineering, model selection, model deployment, and hyperparameter tuning, which are a few of the complicated tasks to make Machine Learning models effective. Additionally, it automated visualisation and Machine Learning interpretation.

3.2.2 Auto-Keras

DATA Lab created the open source software called Auto-Keras [10] library for automatic Machine Learning (AutoML). Auto-Keras offers tools for the automatic search and selection of deep learning model architecture and hyper-parameters. It is simple to use since it adheres to the traditional Scikit-Learn API architecture. The latest version can do deep learning while automatically looking for hyperparameters. In Auto-Keras, the trend is to automate Neural Architecture Search (NAS) techniques to simplify Machine Learning. NAS employs a series of algorithms that change models automatically in place of deep learning practitioners and engineers

3.2.3 TPOT

TPOT [11] is a fully AutoML model and it is positioned as a platform to simplify the regular Machine Learning processes. A genetic algorithm is employed to identify the best model. The greatest predicted accuracy across a wide range of models is being chosen. This framework is a scikit-learn add-on, similar to Auto-Sklearn. However, TPOT follows its algorithms for classification and regression

3.2.4 Dataiku

Dataiku is a well-designed framework to perform automation majority of Machine Learning phases without having any prior programming or Machine Learning experience. This AutoML model quickly and accurately constructs prediction models. Machine Learning models may be easily created with Dataiku's user interface. A business may quickly implement a real-time predictive analytics solution that is powered by a precise Machine Learning model. The ability to dive deeper into the platform and take charge of the Machine Learning workflow is a huge benefit of Dataiku; on the one hand, business analysts can use it as a tool, and on the other hand, skilled data scientists can tune many parameters on their own to get even more accurate models

3.2.5 AutoGluon

Developed by Texas A&M University, AutoCluon [12] is an advanced framework for Automated Machine Learning (AutoML). Its primary objective is to streamline the construction and implementation of machine learning models by automating key steps like feature engineering, hyperparameter tuning, and model selection. By leveraging a robust search algorithm, AutoCluon efficiently explores a wide range of potential machine learning pipelines, identifying the most optimal combinations for enhanced performance

3.2.6 Pycaret

PyCaret is a Python library that simplifies the machine learning workflow process by offering a high-level interface. It is an open-source, simple and low-code solution that automates various steps in the process of building and deploying machine learning models. PyCaret's main goal is to make machine learning accessible to individuals with varying levels of experience, as it significantly reduces the amount of code and effort needed for the implementation of work.

3.2.7 Auto-ML framework comparison

Various Auto-ML frameworks were analysed with respect to basic working principles and supported Machine Learning algorithms (Table 1).

Table 1 Various Auto-ML Frameworks Methodology

AUTOML FRAMEWORK	WORKING METHOD	SUPPORT OF MACHINE LEARNING ALGORITHMS
H2O	H2O returns the best performance from the leaderboard of all models based on the training of two stacked ensembles.	Support Vector Machine, XGBoost, Stacked Ensembles, Naïve Bayes, Gradient Boosting Machine, Generalized Linear Model, Random Forest, and Deep Learning
AutoKeras	Neural Architecture Search (NAS) is a method that makes use of the Keras API and searches over neural network architectures to determine which one would best handle a modelling problem.	Deep Learning and simple Machine Learning models
AutoGluon	Trains a group of models under various conditions and parameters, then chooses the most effective ones by optimising hyper-parameters. A random search, grid search, or Bayesian optimisation is the basis for the search method for the ideal collection of parameters.	K-Nearest Neighbour, Light Gradient Boost Machine, Random Forest, XGBoost, Extremely Randomized Trees, Neural Networks
Pycaret	Pycaret follows step by step process starting from data preparation, model training, hyperparameter tuning, model analysis, model selection, and best results of all models.	XGBoost, Light Gradient Boost Machine, Gradient Boost Classifier, AdaBoost Classifier, Random Forest, Extra Trees, Logistic Regression, K-Nearest Neighbour, Support Vector Machine, Naïve Bayes, and others
TPOT	TPOT used genetic algorithms to automate the design of Machine Learning models and optimize the Machine Learning pipeline.	XGBoost, Decision Tree, logistic regression, random forest, and KNN
Dataiku	Dataiku makes Machine Learning accessible by designing fully automated models in the form of feature extraction, feature selection, and model selection. It enables the selection of models and hyperparameters either manual or automatic.	XGBoost, Decision Tree, Light Gradient Boost Machine, Gradient Boost Classifier, AdaBoost Classifier, Random Forest, Extra Trees, Logistic Regression, K-Nearest Neighbour, Naïve Bayes, Single Layer Perceptron, Support Vector Machine, and others

4. Results and Discussion

4.1 Results

Each AutoML framework is applied to the dataset to evaluate the prediction results. Framework results are shown in Table 2 along with the particular best-performed Machine Learning algorithm. These are aligned with the result which is given by traditional Machine Learning models such as XGBoost. Accuracy is one of the evaluation metrics in Machine Learning. It means the correct prediction rate by the model.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{1}$$

Where TP means True Positive, TN means True Negative, FP means False Positive, and FN means False Negative.

Table 2 Accuracy of AutoML Frameworks

S.NO	AUTOML FRAMEWORK	ACCURACY
1	H2O	92.0% by Stacked Ensemble
2	AutoKeras	56.4% by Deep Learning
3	AutoGluon	92.0% by Random Forest with Gini index
4	Pycaret	91.1% by XGBoost
5	TPOT	72.6% by Random Forest Classifier
6	Dataiku	91.1% by Light Gradient Boost Machine and Gradient Boosting Tree

4.2 Discussion

H2O has produced an accuracy of 92.0% by the Stacked Ensemble model. An accuracy of 92.0% was also given by the Random Forest Classifier with the Gini index of AutoGluon. Both Pycaret and DataRobot have given the same accuracy of 91.1%. Those are provided with various facilities like a selection of particular ML algorithms to be part of prediction to make a better analysis of predicted results. The Random Classifier of TPOT has given an accuracy of 72.6%. Deep Learning of AutoKeras has given an accuracy of 56.4%. Figure 3 shows that algorithms of AutoGluon and H2O give better accuracy over remaining AutoML frameworks.

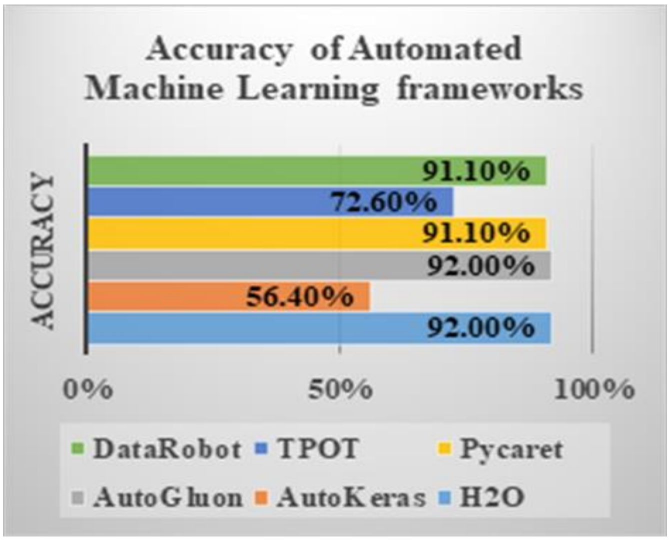

Figure 3 Accuracy of Automated Machine Learning Frameworks

Conclusion

The process of recommending a crop using Machine Learning can be complex and time-consuming, requiring several steps from data collection to the final model prediction result. To overcome this, the incorporation of AutoML, this process can be simplified into a more straightforward, efficient, and effective method. AutoML frameworks are designed to automate the selection of algorithms and features, as well as tune hyperparameters, to produce accurate crop recommendations based on soil parameters. By leveraging AutoML in agriculture, farmers can make more informed decisions regarding crop selection, resulting in higher yields and profits. Overall, AutoML has the potential to revolutionize the agricultural industry, making it automated, more efficient, sustainable, and profitable for all involved.

Acknowledgements

The authors declare that authors have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1]. Yuda N Mnyawami, Hellen H Maziku & Joseph C Mushi Comparative Study of AutoML Approach, Conventional

Ensemble Learning Method, and KNearest Oracle-AutoML Model for Predicting Student Dropouts in Sub-Saharan African Countries, Applied Artificial Intelligence, 36:1, (2022) DOI: 10.1080/08839514.2022.2145632

[2]. Li K-Y, Sampaio de Lima R, Burnside NG, Vahtmäe E, Kutser T, Sepp K, Cabral Pinheiro VH, Yang M-D, Vain A, Sepp K. Toward Automated Machine Learning-Based Hyperspectral Image Analysis in Crop Yield and Biomass Estimation. Remote Sensing. (2022); 14(5):1114. https://doi.org/10.3390/rs14051114

[3]. Dilmurat K, Sagan V, Maimaitijiang M, Moose S, Fritschi FB. Estimating Crop Seed Composition Using Machine Learning from Multisensory UAV Data. Remote Sensing. (2022); 14(19):4786. https://doi.org/10.3390/rs14194786

[4]. Espejo-Garcia B, Malounas I, Vali E, Fountas S. Testing the Suitability of Automated Machine Learning for Weeds Identification. AI. (2021); 2(1):34-47. https://doi.org/10.3390/ai2010004

[5]. Alsharef, A., Aggarwal, K., Sonia et al. Review of ML and AutoML Solutions to Forecast Time-Series Data. Arch Computat Methods Eng 29, 5297–5311 (2022). https://doi.org/10.1007/s11831-022-09765-0

[6]. Xuetao Jiang, Binbin Yong, Soheila Garshasbi, Jun Shen, Meiyu Jiang, Qingguo Zhou. Crop and weed classification based on AutoML[J]. Applied Computing and Intelligence, (2021), 1(1): 46-60. doi 10.3934/aci.2021003

[7]. https://soilhealth.dac.gov.in

[8]. Quanming, Yao, et al. "Taking human out of learning applications: A survey on automated machine learning." arXiv preprint arXiv:1810.13306 (2018): 34.

[9]. LeDell, Erin. "H2O AutoML: Scalable Automatic Machine Learning." (2020).

[10]. Haifeng Jin, Qingquan Song, and Xia Hu. Auto-Keras: An Efficient Neural Architecture Search System. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19). Association for Computing Machinery, New York, NY, USA, (2019)1946–1956. https://doi.org/10.1145/3292500.3330648

[11]. Olson, Randal S. and Jason H. Moore. "TPOT: A Tree-based Pipeline Optimization Tool for Automating Machine Learning." AutoML@ICML (2016).

[12]. Erickson, Nick & Mueller, Jonas & Shirkov, Alexander & Zhang, Hang & Larroy, Pedro & Li, Mu & Smola, Alexander. AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data. (2020).