**ENHANCING H2O AUTOML WITH MAMDANI-TYPE FUZZY INFERENCE SYSTEM FOR SEPSIS PREDICTION**

**MIDTERM REPORT**

**Prepared by:**

**Ahmet Emre Usta**

Student ID: 2200765036

Department of Artificial Intelligence

Hacettepe University

**Course:**

**AIN 421 Fuzzy Logic**

**Instructor:** Prof. Dr. Ebru Akcapinar Sezer

**Submission Date:**

**December 10, 2024**

## Table of Contents

# 1. INTRODUCTION

## Definition of AutoML

Automated Machine Learning (AutoML) is a transformative approach in the field of machine learning that automates the end-to-end process of applying machine learning to real-world problems. By automating tasks such as data preprocessing, feature selection, model selection, hyperparameter tuning, and model evaluation, AutoML democratizes access to machine learning, allowing individuals with limited expertise to build effective predictive models. Examples of tasks automated by AutoML include selecting the best algorithm for a given dataset, optimizing hyperparameters to enhance model performance, and constructing feature engineering pipelines to improve data quality.

## Importance of AutoML

The necessity of AutoML stems from the complexity and time-consuming nature of traditional machine learning workflows. AutoML enhances productivity by significantly reducing the time required to develop models, minimizes human error, and often identifies superior model configurations through exhaustive and systematic search methods. This democratization of machine learning allows non-experts to leverage advanced analytical tools, thereby accelerating innovation and enabling organizations to respond swiftly to data-driven challenges. Additionally, AutoML improves efficiency for experts by automating repetitive tasks, allowing them to focus on more strategic aspects of model development and deployment.

## Project Objective

This project aims to enhance H2O AutoML by integrating a Mamdani-type Fuzzy Inference System (FIS). The integration seeks to leverage fuzzy logic to handle uncertainty and incorporate domain-specific knowledge, thereby improving the model's interpretability and predictive performance in sepsis detection.

## 2. AUTOML SOLUTIONS

**Auto-sklearn**

**Source and Developers:**

Auto-sklearn is an open-source AutoML tool developed by the Machine Learning Research Group at the University of Freiburg, Germany. It is available on GitHub.

**Capabilities:**

Auto-sklearn automates the process of model selection and hyperparameter tuning using Bayesian optimization. It leverages meta-learning to utilize past performance data, enabling it to make informed decisions about which models and configurations to explore. Additionally, it constructs ensembles of multiple models to enhance predictive performance and incorporates feature preprocessing and selection.

**ML Process Execution:**

Users provide datasets in scikit-learn format. Auto-sklearn searches for the best-performing machine learning pipeline by evaluating different models and hyperparameter settings, ultimately returning the optimal model with its configuration. The process involves splitting the data into training and validation sets and performing cross-validation to ensure robust model evaluation.

**TPOT (Tree-based Pipeline Optimization Tool)**

**Source and Developers:**

TPOT is an open-source AutoML tool developed by the Epistasis Lab at the University of Pennsylvania. It is accessible via GitHub.

**Capabilities:**

TPOT employs genetic programming to evolve machine learning pipelines. It automates the selection of models, preprocessing steps, and hyperparameter tuning, allowing for extensive customization through user-defined operators. TPOT can perform automated feature selection and transformation, making it a flexible tool for optimizing machine learning workflows.

**ML Process Execution:**

Users supply datasets compatible with scikit-learn. TPOT evolves pipelines over multiple generations, optimizing for performance metrics such as accuracy. The evolutionary process involves selecting, crossing, and mutating pipeline configurations to discover the best-performing pipeline. The final output is the best pipeline configuration discovered during the evolutionary process.

**H2O AutoML**

**Source and Developers:**

H2O AutoML is an open-source AutoML solution developed by H2O.ai. It is available on GitHub.

**Capabilities:**

H2O AutoML supports a broad range of algorithms, including Generalized Linear Models (GLM), Gradient Boosting Machines (GBM), Random Forest, Deep Learning, and Stacked Ensembles. It is designed for scalability, efficiently handling large datasets and complex models. H2O AutoML also facilitates automatic stacking, where multiple models are combined to improve predictive performance. Additionally, it offers multi-language support, allowing users to interact with the tool via Python, R, or the H2O Flow GUI.

**ML Process Execution:**

Users can initiate AutoML processes via Python, R, or the H2O Flow GUI. H2O AutoML automates data preprocessing, model training, hyperparameter tuning, and ensemble building. It outputs a leaderboard ranking models based on performance metrics, along with the best-performing model and its ensemble. The process is designed to be user-friendly, accommodating both small and large-scale datasets.

**Comparative Analysis**

FeatureAuto-sklearnTPOTH2O AutoML

**Source** Open-source Open-source Open-source with enterprise options

**Developers** University of Freiburg University of Pennsylvania H2O.ai

**Model Support** scikit-learn models scikit-learn models Wide range including deep learning

**Optimization Method** Bayesian Optimization Genetic Programming Automated Stacking and Hyperparameter Optimization

**Ease of Use** Python-centric integrates with scikit-learn Python-centric, code-based Supports multiple languages, GUI available

**Scalability** Suitable for small to medium datasets Suitable for small to medium datasets Highly scalable for large datasets

**Community Support** Active community, frequent updates Active community, frequent updates Strong support from H2O.ai and community

**Extensibility** Limited customization High customization via genetic operators Moderate customization, primarily via API

**Strengths of H2O AutoML:**

- **Scalability:** Capable of handling large datasets efficiently.
- **Comprehensive Model Support:** Includes advanced algorithms like deep learning.
- **Integration and Deployment:** Supports multiple languages and provides robust deployment options.
- **Automatic Stacking:** Enhances predictive performance through ensemble methods.
- **Active Development and Support:** Backed by H2O.ai and ensures continuous updates and strong community backing.

**Weaknesses of Auto-sklearn and TPOT:**

- **Auto-sklearn:** Limited scalability compared to H2O AutoML and narrower model support.
- **TPOT:** While highly customizable, it also faces scalability limitations and focuses primarily on scikit-learn models.

## 3. SELECTED PLATFORM AND EXTENSION DESIGN

### Chosen AutoML Tool: H2O AutoML

### Rationale for Selection

After evaluating Auto-sklearn, TPOT, and H2O AutoML based on criteria such as scalability, model support, integration capabilities, and community support, H2O AutoML emerged as the most suitable choice for enhancement. Its ability to handle large datasets, support for a wide array of models including deep learning, and robust deployment options make it ideal for real-world applications. Additionally, the flexibility of its API allows for seamless integration of custom components like a Fuzzy Inference System.

### Design of FIS Extension

### FIS Overview

A Mamdani-type Fuzzy Inference System (FIS) is chosen for its interpretability and rule-based approach, which aligns well with the need to incorporate domain-specific knowledge into the machine learning pipeline. The FIS will handle uncertainties in critical features, providing a refined input for the AutoML process.

### Integration Strategy

The FIS will be integrated as a preprocessing step within the H2O AutoML pipeline. Specifically, it will:

1. **Enhance Feature Engineering:**
    - Apply FIS to selected features to create additional fuzzy-based features that capture nuanced relationships.
2. **Incorporate Domain Knowledge:**
    - Utilize fuzzy rules derived from medical expertise to inform the machine learning models, improving interpretability and decision-making.

**Technical Specifications**

- **Membership Functions:**
  - Triangular membership functions will be used for simplicity and ease of interpretation.
- **Fuzzy Rules:**
  - A comprehensive set of if-then rules will be defined based on domain knowledge in sepsis prediction. For example:
    - *IF heart_rate is High AND temp is High THEN risk is High.*
    - *IF resp_rate is Low AND wbc is Low THEN risk is Low.*
- **Defuzzification Method:**
  - The Center of Area (COA) method will be employed to convert fuzzy outputs into crisp values suitable for integration with H2O AutoML.

**Implementation Plan**

1. **Identify Key Features:**
   - Select vital features relevant to sepsis prediction (e.g., heart rate, respiratory rate, temperature).
2. **Define Fuzzy Variables and Membership Functions:**
   - Create linguistic variables with triangular membership functions for each selected feature.
3. **Develop Fuzzy Rules:**
   - Establish a comprehensive rule set capturing the relationships between input features and the sepsis risk.

4. **Implement FIS:**
    - Utilize the scikit-fuzzy library in Python to build the Mamdani-type FIS.
5. **Integrate FIS with H2O AutoML:**
    - Develop a preprocessing module that applies FIS to the dataset, generating additional features or refining existing ones.
    - Incorporate this module into the H2O AutoML pipeline to ensure seamless data flow.
6. **Testing and Validation:**
    - Compare the performance of the enhanced H2O AutoML pipeline with the standard pipeline using evaluation metrics such as accuracy, ROC AUC, and F1 Score.

## 4. CONCLUSION

### Summary of Findings

This project reviewed top AutoML solutions—Auto-sklearn, TPOT, and H2O AutoML—and selected H2O AutoML based on comprehensive criteria including scalability, model support, integration capabilities, and community support. The design approach involves enhancing H2O AutoML with a Mamdani-type Fuzzy Inference System (FIS) to incorporate domain-specific knowledge and handle uncertainties in data, aiming to improve both model interpretability and predictive performance in sepsis detection.

### Expected Outcomes

- **Improved Handling of Uncertainty:** The integration of FIS will enable better management of uncertainty in critical features through fuzzy logic.
- **Enhanced Model Interpretability:** Fuzzy rules will provide clear, interpretable insights into how input features influence predictions.

- **Performance Gains:** The enhanced AutoML pipeline is expected to achieve higher accuracy and better overall performance metrics in sepsis prediction models.

**Future Work**

- **Implementation of FIS Extension:** Proceed to develop and integrate the FIS extension with H2O AutoML.
- **Comprehensive Testing:** Conduct thorough testing and evaluation of the enhanced pipeline against the standard H2O AutoML to quantify performance improvements.
- **Deployment and Monitoring:** Explore deployment strategies for the enhanced model and establish monitoring mechanisms to ensure sustained performance in production environments.

## 5. REFERENCES


- Auto-sklearn GitHub Repository: https://github.com/automl/auto-sklearn
- TPOT GitHub Repository: https://github.com/EpistasisLab/tpot
- H2O AutoML GitHub Repository: https://github.com/h2oai/h2o-3
- H2O.ai Official Documentation: https://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html
- Scikit-Fuzzy Documentation: https://scikit-fuzzy.github.io/scikit-fuzzy/