Ensemble Diversity in SageMaker Autopilot: Robustness and Accuracy Analysis

Assignee Research

June 12, 2026

Abstract

Abstract Feature selection becomes prominent, especially in the data sets with many variables and features. It will eliminate unimportant variables and improve the accuracy as well as the performance of classification. Random Forest has emerged as a quite useful algorithm that can handle the feature selection issue even with a higher number of variables. In this paper, we use three popular datasets with a higher number of variables (Bank Marketing, Car Evaluation Database, Human Activity Recognition Using Smartphones) to conduct the experiment. There are four main reasons why feature selection

1 Introduction

This paper examines: Selecting critical features for data classification based on machine learning methods. Research question: To what extent does the ensemble diversity in SageMaker Autopilot affect its robustness and accuracy compared to single-model AutoML solutions like H2O.ai and TPOT on the Amazon Employee Access dataset?.

2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.5/10.

3 Results

11 papers retrieved. 12 claims extracted; 9 independently verified. Quality review score: 7.5/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
Feature selection eliminates unimportant variables and improves the accuracy as well as the performance of classificatio	✓	0.19
Random Forest can handle the feature selection issue even with a higher number of variables.	✓	0.29
The study uses three datasets: Bank Marketing, Car Evaluation Database, and Human Activity Recognition Using Smartphones	✓	0.22
The three datasets used in the experiment have a higher number of variables.	✓	0.16
Feature selection simplifies the model by reducing the number of parameters.	✓	0.17
Feature selection decreases training time.	×	0.10
Feature selection reduces overfitting by enhancing generalization.	×	0.10
Feature selection helps avoid the curse of dimensionality.	×	0.14
The study evaluates and compares the accuracy and performance of Random Forest (RF), Support Vector Machines (SVM), K-Ne	✓	0.31
The paper adopts Random Forest to select important features in classification.	✓	0.24
The study compares results with and without feature selection using RF methods varImp(), Boruta, and Recursive Feature E	✓	0.22
The study measures best percentage accuracy and kappa to evaluate feature selection methods.	✓	0.17

References

- https://doi.org/10.63125/edxgjg56

- https://doi.org/10.1186/s40537-020-00327-4
- https://doi.org/10.1109/access.2022.3207287