

## How to make automatic feature discovery efficient and effective?



### Efficient & Effective

How to make automatic feature discovery efficient and effective? Efficient means fast and scalable. Effective means accurate and robust.

How to make automatic feature discovery efficient and effective? Efficient means fast and scalable. Effective means accurate and robust.

How to make automatic feature discovery efficient and effective? Efficient means fast and scalable. Effective means accurate and robust.

How to make automatic feature discovery efficient and effective? Efficient means fast and scalable. Effective means accurate and robust.

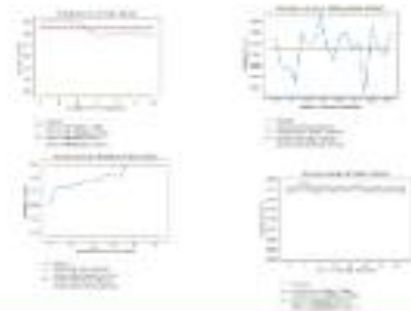
Dataset	Time	F1 Score	Robustness
Dataset 1	10 min	0.85	High
Dataset 2	15 min	0.80	Medium
Dataset 3	20 min	0.75	Low
Dataset 4	25 min	0.70	Very Low

## Experiments

Our experiments are performed on 4 different real-world datasets. Best to worst and ranked by accuracy.

Dataset	Method	Time	F1 Score	Robustness
Dataset 1	Heuristic	10 min	0.85	High
Dataset 2	Machine Learning	15 min	0.80	Medium
Dataset 3	Heuristic	20 min	0.75	Low
Dataset 4	Machine Learning	25 min	0.70	Very Low

## Main Results



## Methodology

Our methodology consists of several steps. The main idea is to discover features by analyzing the characteristics of the data and selecting the most relevant ones.

1. Feature Selection
2. Feature Engineering
3. Feature Evaluation
4. Feature Refinement
5. Feature Ranking
6. Feature Selection
7. Feature Engineering
8. Feature Evaluation
9. Feature Refinement
10. Feature Ranking



Our methodology is a heuristic approach to feature discovery. It is fast and scalable.

Our methodology is a heuristic approach to feature discovery. It is fast and scalable.



Our methodology is a heuristic approach to feature discovery. It is fast and scalable.

Our methodology is a heuristic approach to feature discovery. It is fast and scalable.



## Conclusion

The heuristic performs well in 3/4 datasets. It performs equally or better than all other experiments, with very K values less than 5.

## Future work

Collect more data on the datasets for higher precision or further analysis. A collection of data thousands of datasets instead of hundreds could allow for better analysis of the covariance between statistical characteristic scores and the independence of the likelihood.

Pair with machine learning:

The heuristic itself could be fine-tuned using a reinforcement learning algorithm, or possibly used as an input to train a machine learning model that predicts the utility of the feature.

Pair with other heuristics:

Pairing the model with heuristics that take into account the statistical values rather than just the rankings might synergize well with a heuristic that only considers ranking of features, as it is the heuristic I developed is less precise when all features similar statistical characteristics.