# Automatic Feature Discovery: A comparative study between filter and wrapper feature selection techniques

Andrei Mânăstireanu a.b.manastireanu@student.tudelft.nl
Responsible Professor: Asterios Katsifodimos a.katsifodimos@tudelft.nl
Supervisor: Andra Ionescu a.ionescu-3@tudelft.nl

## Research Question

*How do different feature selection techniques for categorical and numerical data influence the performance of simple decision trees, linear machine learning algorithms and support vector machines?*

## Preliminaries

*Feature selection* reduces dimensionality by selecting a subset of relevant features [1].

*Filter techniques* use statistical tests or other mathematical calculations [1], while *wrapper techniques* use a machine learning model [1].

Filter methods: *Chi-Squared*, *ANOVA*.
Wrapper methods: *Forward Selection*, *Backward Elimination*.

Machine Learning models: *Gradient Boosting Machine* (GBM), *Extreme Gradient Boosting* (XGB), *Random Forest* (RF), *Linear Regression*, *Logistic Regression*, *Support Vector Classification* (SVC), *Support Vector Regression* (SVR).

## Results

The percentage of selected features varies between 0% and 100%. The runtime is expressed in seconds.

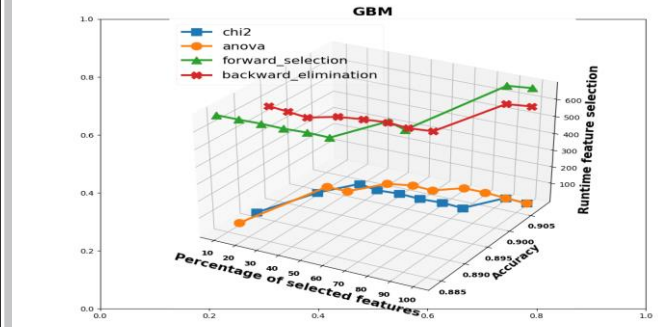


Figure 1: Accuracy of Gradient Boosting Machine for the bank marketing dataset (experiment 2)

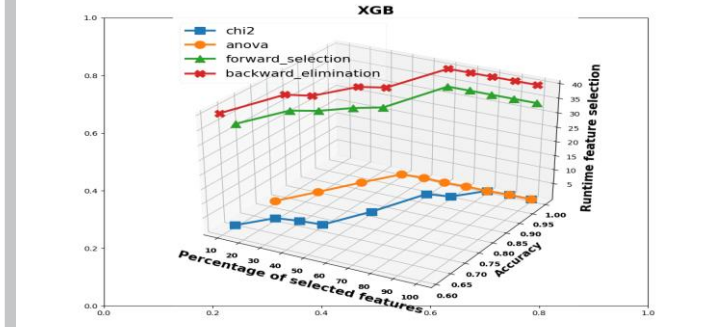


Figure 3: Accuracy of Extreme Gradient Boosting for the steel plates faults discrete subset (experiment 4)

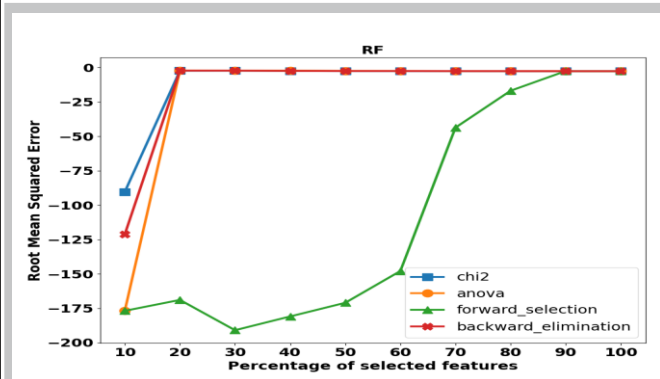


Figure 2: Root Mean Squared Error of Random Forest for the bike sharing dataset (experiment 2)

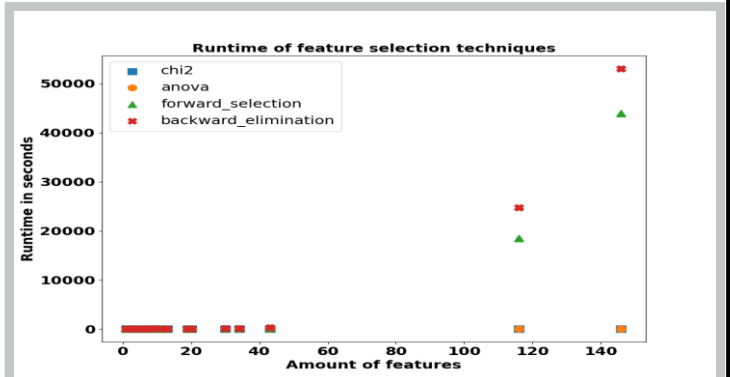


Figure 4: Runtime of feature selection techniques w.r.t. number of features of each dataset (experiment 4)

## Conclusions and Future Work

*Filter* methods *outperform wrapper* methods regarding classification *accuracy*, *regression root mean squared* error, and *runtime*.

The analysis of dataset structure in terms of categorical and numerical features shows that:

- *Chi-Squared* and *ANOVA* are particularly suitable for *categorical* data.
- *ANOVA* performs better for *continuous* numerical data.
- In the case of *discrete* numerical data, *Chi-Squared* and *ANOVA* should be used.

*Future work* can expand the collection and analysis of datasets and investigate the use of alternative underlying estimators for wrapper methods.

## References

[1] - Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. ACM computing surveys (CSUR), 50(6), 1-45.