**AN INVESTIGATION OF FEATURE SELECTION METHODS FOR CLASSIFICATION PURPOSES USING SUPERVISED MACHINE LEARNING**

**M.Sc., BUSE NUR BALTACIOĞLU1, Assoc. Prof. Dr. İDİL YAVUZ2**

**1**Dokuz Eylül University, The Graduate School of Natural and Applied Sciences,

ORCID: 0000-0001-9848-2332

**2**Dokuz Eylül University, Faculty of Science, ORCID: 0000-0003-2163-1066

**ABSTRACT**

With recent developments in computer technologies, data from all sciences have become easier to access and store. Bigger data sets are available in almost all fields and this as a result gave rise to the machine learning algorithms for knowledge discovery since they can help resolve many challenges classical statistical approaches face under these circumstances. One of these challenges is the dimensionality problem where there are many features but not as many observations to provide dependable estimation of model parameters for all features. Even though with machine learning we switch our focus to better estimates rather than interpretability, still a trade-off is needed. Therefore, feature selection is still a crucial step in learning since proper selection of features helps reduce possible biases, computational complexity and loss of interpretability. The literature on feature selection is initially categorized into three subgroups, namely: Filter, Wrapper and Embedded approaches. In this study we discuss and compare frequently used Filter methods, namely information gain, gain ratio, symmetrical uncertainty, chi-squared test, Cramer’s V and OneR algorithm. We also provide an application for the detection of early stage diabetes using logistic regression. We compare the predictive powers of the models built using the famous filtering approaches by comparing the test and cross validation accuracies. As a result, it can be seen that the filtering approaches produce similar predictive accuracy when very few or many features are filtered however the OneR algorithm provides better accuracy when a more balanced number of features are filtered.

**Keywords :** Supervised machine learning, feature selection, filter methods, OneR algorithm