TRANSACTIONS ON EMERGING SCIENCES
Ideas at the Edge of Discovery
ISSN 3136-2710

ORIGINAL RESEARCH ARTICLE

ANALYSIS OF FILTER-BASED FEATURE SELECTION METHODS USING TOP FEATURE SUBSETS AND MULTIPLE EVALUATION METRICS

Ayesha Javed
Department of Computer Science,
University of Agriculture, Faisalabad
Corresponding Author Email: fatisha843@gmail.com

ABSTRACT

Feature selection is an important step in improving the performance and efficiency of machine learning models, especially when handling high-dimensional datasets. The presence of irrelevant and redundant features can reduce model accuracy and increase computational cost. In this study, several widely used filter-based feature selection techniques are examined, including Chi-Square, Analysis of Variance (ANOVA), Mutual Information, and correlation-based methods. The main objective is to analyze how these techniques perform when selecting different numbers of top-ranked features.To ensure a fair evaluation, multiple datasets with different characteristics are used. For each dataset, feature subsets are created by selecting the top 20, 40, 60, and 80 features based on ranking scores. These subsets are then used to train machine learning models, and their performance is evaluated using metrics such as accuracy, precision, F1-score, and Root Mean Square Error (RMSE). This approach helps in understanding how the number of selected features influences model performance. The results show that feature selection has a significant impact on model outcomes. In many cases, better performance is achieved with a moderate number of features rather than using all available features. It is also observed that the effectiveness of each method varies depending on the dataset. Overall, this study provides useful insights for selecting appropriate feature selection techniques in practical machine learning applications.

Keywords: *Feature Selection, Filter Methods, Chi-Square, ANOVA, Mutual Information, Correlation, Top Features, Machine Learning, Accuracy, F1-Score, RMSE, Precision, MultiDataset Analysis*

1. INTRODUCTION

The extensive availability of information and data within numerous areas has considerably raised the complexity of applications related to machine learning. Current datasets tend to be highly dimensional, meaning that they may contain a relatively large number of features compared to the number of samples (1). It is problematic for such data to be processed because there is always a danger of noise features distorting actual patterns in data (2). As a consequence, feature selection has emerged as one of the vital preprocesses aimed at optimizing the modelling process (3).

As the name implies, feature selection is a practice of choosing a particular set of informative features, which is able to provide a satisfactory representation of the original dataset while ignoring the rest (4). The process helps simplify computational complexity and enhance the generalization power of a model through decreasing the impact of noise (5).

ISSN 3136-2710

There are different methods of performing feature selection, including filter, wrapper, and embedded techniques. Filter methods constitute the most popular among these because of their efficiency and the ability to perform independently of a given modeling algorithm (6). There is several filtering approaches used to measure the importance of each feature in the data set using various criteria. The Chi-Square test estimates the relationship between the categorical features and the output. Analysis of variance measures the differences between group means. Mutual Information detects both linear and non-linear dependencies, whereas correlation-based tests estimate the degree of dependence among features (7). Given the distinct nature of their criteria, the results may differ.

Although there are some benefits associated with feature selection, choosing an efficient method is a difficult task. First, most research articles use either one data set or fix the number of features for their investigation. Furthermore, there is no systematic evaluation that discusses the performance of feature selection based on different data sets and varying numbers of feature subsets. Such an evaluation framework can be a useful contribution in the field.

In addition to finding a proper feature selection algorithm, another key problem concerns the choice of the optimal number of features. Over-selecting features may lead to redundant information; however, under sampling may cause a loss of important features (8). Thus, the analysis of model performance about varying numbers of feature subsets becomes indispensable. In this experiment, feature subsets are created based on the most highly ranked features, namely Top 20, Top 40, Top 60, and Top 80, for evaluating their effect on model performance.

To conduct an accurate assessment of model performance, several metrics have been taken into consideration. Firstly, accuracy indicates model performance in general. Secondly, F1-score accounts for the trade-off between precision and recall. Lastly, Root Mean Square Error (RMSE) is adopted to calculate the difference between predicted and true values.

There are several works focused on comparing various filter-based feature selection algorithms; however, most of these works have a very limited number of datasets used and do not vary in the number of selected features. Consequently, there is no clear answer regarding the best technique. The current study aims to fill this gap and compare the effectiveness of various feature selection approaches on multiple datasets and varying numbers of selected features.

The Objectives of this study are:

- To conduct an evaluation of several feature selection techniques for multiple datasets.
- To analyse how the variation in the size of the feature subsets (Top 20, 40, 60, and 80) affects model performance.
- To estimate model performance based on multiple criteria: accuracy, F1 score, and RMSE.
- To determine the most effective technique for feature selection.

The research represents a comprehensive evaluation of multiple filter-based feature selection approaches within one experimental setting. By analyzing several datasets and using various feature subset sizes, the work will help identify the most appropriate algorithm for use.

2. METHODOLOGY

Afterwards, we take another step further to investigate the effect of different methods of feature selection by different method (ANOVA, Chi-Square, Correlation, Mutual Information) on machine learning by considering the use of random datasets which have been trained by different approaches using different algorithms like Decision Tree method(DT), Neighbor Joining(NB), Support Vector Machine(SVM), Linear regression(LR) , and K -Nearest Neighbor(KNN) to maintain the consistency of all the experiments performed. The following is the flow chart depicting our methodology.

ISSN 3136-2710

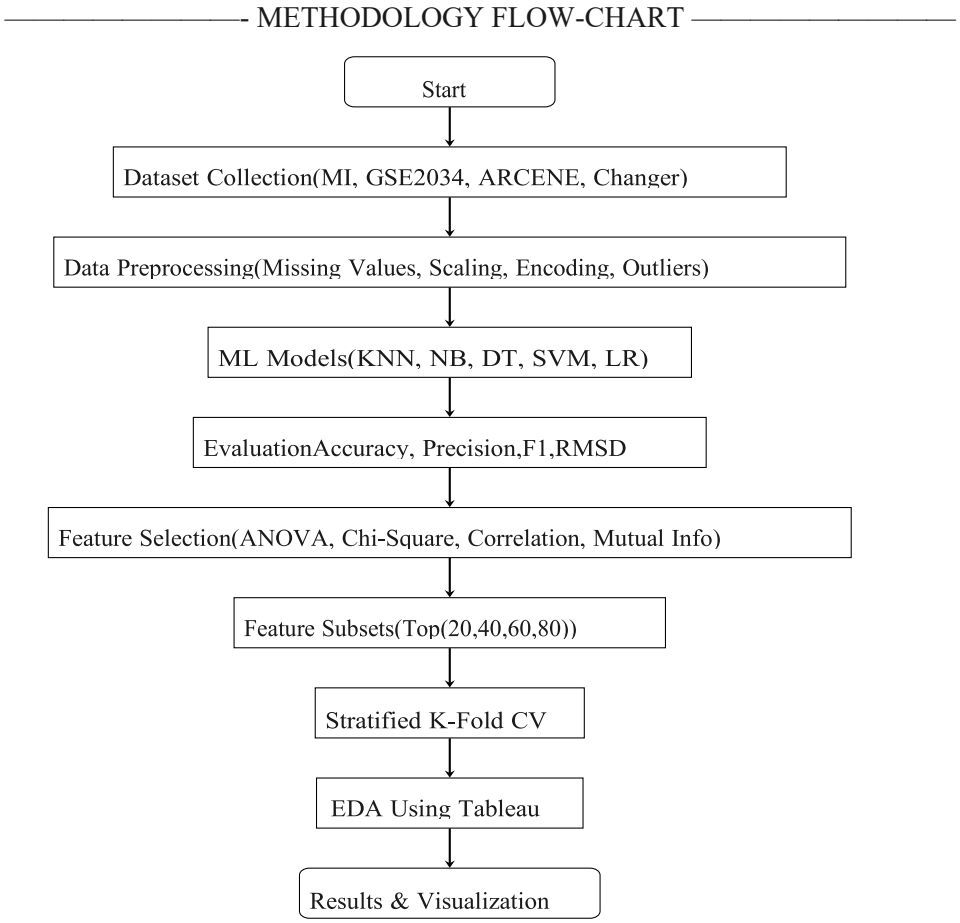


Figure 1: Overall Methodology Flowchart

This study evaluates how different feature selection strategies influence classification outcomes when combined with various machine learning models (3).

Description of Datasets

For evaluating the effectiveness of the proposed method, we use four randomly selected datasets as described below:

- MI Dataset: This dataset consists of high-dimensional features that contain different degrees of redundancy among them. This dataset is generally used for validating the efficiency of selecting significant features and algorithms. We perform our experiment by training this dataset with different fold values and then applying the algorithm to it (9).
- GSE2034 Dataset: GSE2034 is a genetic dataset used extensively in the medical domain for classifying diseases. This dataset is characterized by having a much higher number of features than samples, which makes feature selection critical and interpretable(10).
- ARCENE Dataset: This is another dataset with high dimensionality, which is constructed using mass spectrometry. This dataset also includes irrelevant features and noise, thus making it a good candidate for validating feature selection approaches (11).

ISSN 3136-2710

- Changer Dataset: The Changer benchmark dataset can be used for assessing consistency and robustness of machine learning algorithms using various experiment conditions. The Changer benchmark data set is widely used for testing robustness and consistency of feature selection algorithms and classifiers in varying conditions (12).

Data pre-processing

The data pre-processing step is important in machine learning models because most of the raw data may have problems like missing values, outliers, and inconsistent feature scaling.

Missing value problems are solved by statistical imputation methods such as mean and median imputations based on distribution of the data (13). The feature scaling step involves using standardization so that each feature makes equal contributions in distance-based algorithms like KNN and SVM. Categorical data, if any, are converted to numerical data using encoding. Detection and treatment of outliers help reduce the effects on model training. Several models are used for training datasets using k-fold validation (14).

Feature Selection Approaches

Feature selection is one important aspect of this paper (15). Its objective is the reduction in data dimensionality through the inclusion of the most informative features for accurate classification (16). Four commonly employed approaches to filter-based feature selection are explored in this study.

- ANOVA (Analysis of Variance): ANOVA ranks and selects features based on their statistical significance with respect to the target attribute, by considering the inter-class variance. Inter-class variances for features that have higher values are taken as more significant (17).
- Chi-square Test: This statistical test considers the level of association between the categorical features and the target class. Features with high Chi-square values exhibit strong association with the class label and are ranked as more informative (18).
- Correlation-Based Selection: The Correlation-Based Feature Selection employs Pearson correlation to find the set of most correlated features and eliminate redundancy from the dataset. Features exhibiting low correlation are selected for further processing (19).
- Mutual Information: Mutual information takes into account both linear and non-linear associations between features and classes. Features having high mutual information values are selected and used for further analysis (20; 19).

Having prioritized the features using each technique, the following four sets are formed by taking features from the top of the rank list:

- First 20 features
- First 40 features
- First 60 features
- First 80 features

Each feature set is independently examined under different data partitioning methods, that is, 40/60, 30/70, and 20/80 for training and testing data, respectively.

ISSN 3136-2710

Classification Models and Learning Algorithms

In order to assess the effectiveness of selected feature subsets, five popular machine learning algorithms have been used in this research. The following classifiers are considered because of their different learning methods and ability to classify.

- K-Nearest Neighbors (KNN): KNN is a non-parametric and instance-based classification model that uses majority voting to assign a label based on the class of the nearest data points in the feature space. In order to find similar points, KNN employs a distance function such as the Euclidean distance (21).
- Gaussian Na¨ıve Bayes: Gaussian Na¨ıve Bayes is a probabilistic classification algorithm that uses Bayes' theorem to predict labels. It considers each feature independently to compute the conditional probability. In this regard, features follow the normal distribution. This algorithm works effectively even on large datasets due to less computational cost.
- Decision Tree: Decision Tree is an algorithm that uses supervised learning to divide the training set into smaller sets according to the values of certain features. Decision Tree builds a tree where each node corresponds to certain rules and each leaf corresponds to the target label (22). The model is simple to understand but susceptible to overfitting without proper regulation.
- Support Vector Machine (SVM): Support Vector Machine (SVM) is an excellent classification algorithm that determines the optimal hyperplane for dividing data points belonging to different classes with the maximum possible margin. It works well in highdimensional space and can accommodate non-linear dependencies through kernels such as radial basis function (RBF) (23).
- Logistic Regression: Logistic Regression is a widely used probabilistic model based on statistics that helps predict the probability of occurrence of one or another event using the logistic (sigmoid) function. This method establishes dependence between the input parameters and the target variable, which makes it efficient and understandable (24).
-

Cross-Validation Model Evaluation

To get an unbiased estimate of how the models will perform, Stratified K-Fold Cross-Validation is used (6). This cross-validation method retains the class distribution across all folds, making it appropriate to use for classification tasks.

In this study, cross-validation is performed when evaluating the model after the feature selection process and configuring the model training data. The data are split into K folds for each feature subset (top 20, top 40, top 60, and top 80). During each round, one of the K folds acts as the validation set, while the other $K-1$ folds are used as the training set to train the model. In the end, all the folds become validation sets once. The performance scores from all the iterations are averaged to get the final evaluation result of each model and feature subset (25). Thus, using the cross-validation method helps eliminate random fluctuations when splitting the data.

Cross-validation is performed after the feature selection step to ensure that each model is evaluated equally regardless of the number of features used in a particular combination.

Evaluation Metrics for Performance Measurement

In order to properly evaluate the efficiency of the employed classifiers, several metrics will be used. Every metric will provide a unique insight into the model performance under various conditions of feature selection.

ISSN 3136-2710

- Accuracy: Accuracy is defined as the ratio of correctly classified samples compared to the total number of samples. It gives a general idea about how well a classifier performs. However, its usage is questionable in case of imbalanced datasets where one class significantly outweighs other classes (26).
- Precision: Precision is a metric indicating the ratio of correctly classified positive samples to the total number of positive predictions. This parameter is critical in situations where false-positive predictions carry significant costs (27).
- F1-Score: The F1-score is a statistic reflecting a balance between precision and recall.
This metric is widely used when assessing classification models on imbalanced datasets (28).
- Root Mean Square Error (RMSE): RMSE measures the average size of the error in the predictions made by taking the square root of the average of the squared differences between predicted and observed values. This helps to understand how well the predictions of the model match reality, thus providing information about the accuracy of predictions
(29).

Exploratory Data Analysis (EDA)

Exploratory Data Analysis is carried out using Tableau to comprehend the nature of the datasets.
Visualization methods can assist in recognizing any patterns, correlations, and anomalies in the dataset, which are associated with its top-ranked features.
Line Bar Chart can be utilized to analyze the distribution of the features, plots for outlier detection, Scatter Plot for examining relationships among variables, and Bar Chart for studying the correlation among the features (30).

3. RESULTS

Overview of Experimental Results

This chapter is based on the experimental findings that have been derived by conducting feature selection on various data sets, namely MI, GSE2034,(31), ARCENE, and Changer. First, the data sets were pre-processed to select the best ranked features with the help of feature selection algorithms like ANOVA, Chi-Square, Correlation, and Mutual Information.
Next, these best ranked features were used to train ML models, whose performance was analyzed using various metrics such as Accuracy, Precision, F1-Score, and Root Mean Square Error (RMSE). These experiments have been carried out for comparison purposes by visualizing the results through Tableau.

Visualization of Top Feature Performance

Tableau was employed to depict the model performance when they were trained using topranked features. These visualizations have been made in the form of comparison plots for accuracy, F1 score, precision, and RMSE for each feature selection approach on all data sets. These visualizations give insight into how feature selection affects classification performance. Since all other features were stripped out, only top-ranked features were considered, allowing for better performance most of the timeas shown in fig 2.
As shown that CD3 and GSE2034 datasets have shown to be quite consistent in their performance levels despite the different algorithms being used in their modeling. On the other hand, the consistency of ARCENE and MI is quite low; therefore, the application of a different algorithms will greatly influence the results.

ISSN 3136-2710

Figure 2: Comparison of different datasets with different paramter

Evaluation Metric Comparison

Model performances were compared using the following evaluation metrics:

Accuracy: As per the result, it can be concluded that models trained using selected top features have high accuracy compared to when all features were used. Therefore, it can be said that eliminating redundant features helps improve classification.

Precision: From the results, it can be concluded that precision also improved after applying feature selection to most of the datasets because the models generated fewer false positives.

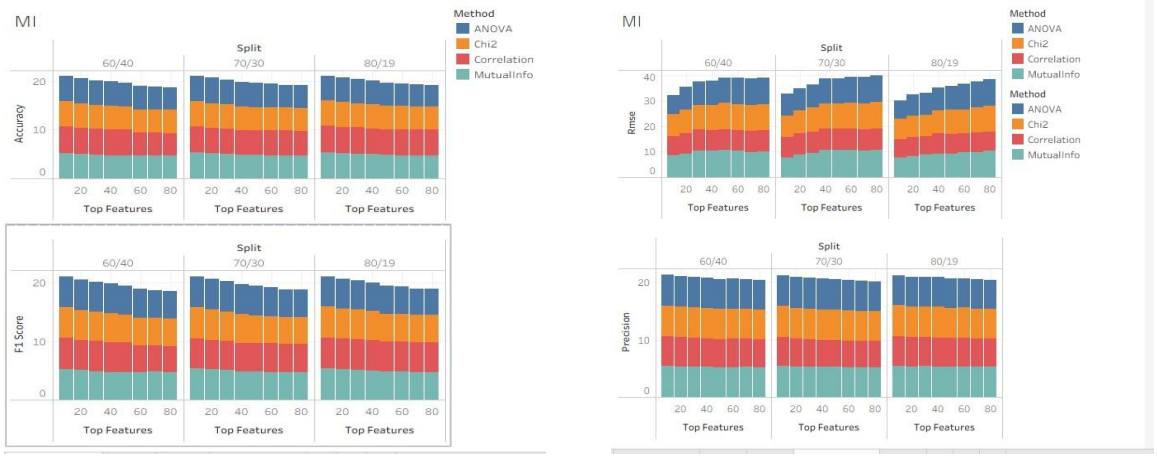
F1-Score: From the results, it can be seen that most of the datasets saw a consistent increase in F1 score because it signifies better precision and recall.

RMSE: Results from this metric clearly depict a decrease in RMSE after feature selection and model become more reliable.

ISSN 3136-2710

Dataset-wise Performance Analysis

The MI dataset exhibited better results in terms of all performance indicators after feature selection based on their accuracy, F1 score, RMSD value, and precision, suggesting that there were redundant features in the initial dataset.as shown in fig3a and 3b.



(a) Top feature accuracy of MI

(b) RMSD of top features

Figure 3: Visual representation of MI top Features

In the case ofGSE2034 data set, the feature selection greatly improved accuracy and F1score since the data were highly high-dimensional. This shows the necessity of applying dimensionality reduction in gene expression for the datasets using line chart with different symbol representations as shown in fig 4.

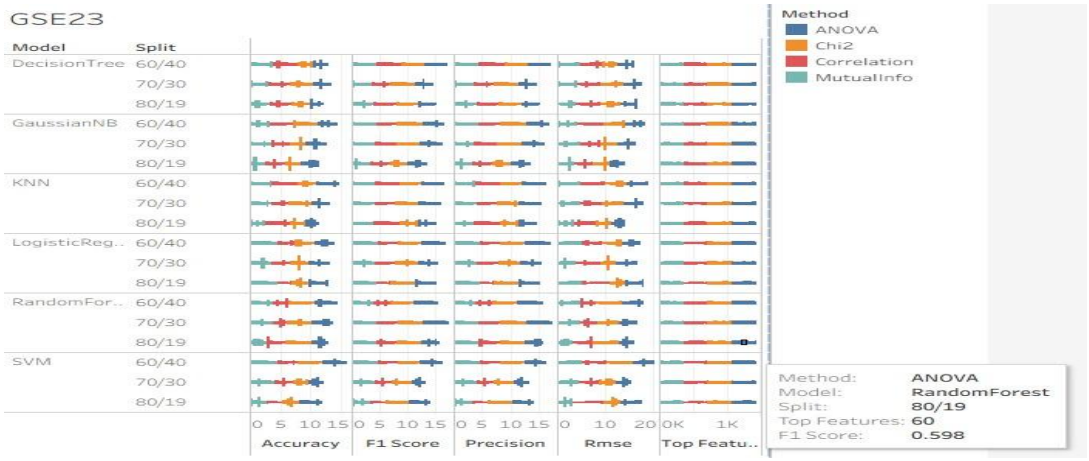


Figure 4: GSE top feature analysis visualization

The ARCENE dataset demonstrated the most pronounced performance enhancement after feature selection. Specifically, the performance enhancements were significant on the basis of the F1-score and RMSE measures. This may be attributed to the fact that this dataset has many noisy features, which can be observed by employing a horizontal line chart with square-shaped boxes.as in fig 5.

ISSN 3136-2710

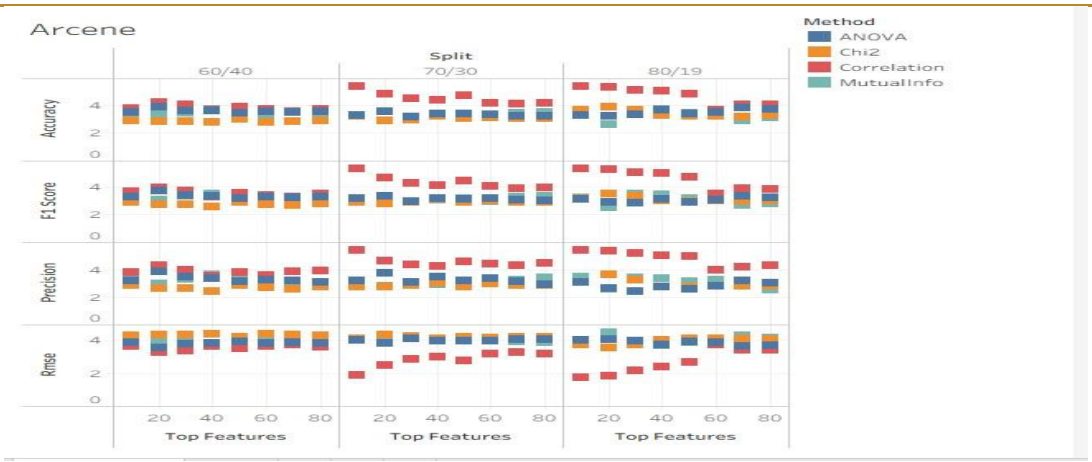
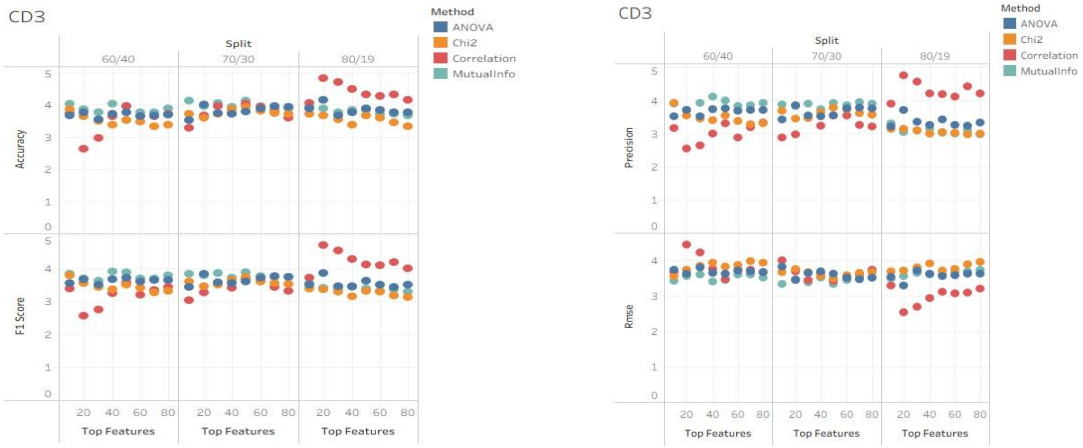


Figure 5: Arcene generalization of top features

The dataset called Changer had quite stable behavior with any feature selection method that was used, which proves that there is a good balance between features in this dataset.6a and 6b.



(a) Top feature accuracy of Changer dataset (b) top features precision of Changer

Figure 6: Visual representation of Changer top Features

In comparison to all other feature selection algorithms, mutual information and ANOVA achieved more accurate outcomes than chi-square and correlation methods. The method of mutual information exhibited excellent results since mutual information is capable of modeling not only linear but also nonlinear connections between variables.

The method of ANOVA also achieved good results where separation between classes was evident. The results of using chi-square were average, whereas correlation-based selection assisted in avoiding redundancy.

Summary of Results

In general, the findings show that there is an increase in performance when high-ranking features are selected. It is seen that the use of good feature selection techniques and the incorporation of machine learning models

ISSN 3136-2710

have greatly helped in achieving increased classification accuracy, precision, F1-scores, and reduced RMSEs. This can be clearly seen through the visualizations done by using Tableau.

4. DISCUSSION

In this section, an elaborate discussion on the findings gained from employing various feature selection approaches and models on several datasets is presented. From the results shown above, it can be noted that feature selection has a considerable effect on improving model performance. The use of feature selection led to the attainment of better Accuracy and F1-score while lowering the RMSE because irrelevant and redundant features were omitted. As far as the used feature selection approaches go, Mutual Information, and ANOVA exhibited superior performance in comparison with Chi-Square and correlation approaches. Mutual Information proved to be a better method owing to its ability to detect nonlinear relationships among features (19). ANOVA performed well on datasets with distinct class boundaries. Chi-Square performed moderately, especially in relation to categorical features, while Correlation-Based Feature Selection aided in eliminating redundant variables.

From the analysis conducted on different levels of feature selections (20%, 40%, 60%, and 80%), it is evident that model accuracy does not always increase as the number of selected features increases. Many times, the best accuracies were achieved using 40% and 60% levels of feature selection. Using 20%, the system may lack important information. The use of 80% level may still introduce redundant variables which negatively affect the performance of the models. As such, moderate subset of features gives better results.

From the analysis, Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) models performed significantly improved their performance due to feature selection process. SVM performed excellently since the model handles well high dimensional spaces (23). KNN model showed good results since the introduction of noise was minimal. Decision tree model performed moderately but was highly vulnerable to overfitting. Gaussian naïve bayes showed consistent performance but performed poorly due to dependency assumptions. Logistic regression performed well but poorly with complex relationships.

The MI dataset experienced improvement through feature selection as a result of the presence of redundant features. The GSE2034 dataset gained the most benefit from feature selection due to the high-dimensionality of the dataset. The ARCENE dataset gained the highest improvement by eliminating noisy features. The Changer dataset exhibited stable results across all feature selections.

It can be concluded that feature selection plays a vital role in improving the performance of classification algorithms when dealing with high-dimensional data. The use of Mutual Information along with SVM and a moderate level of feature selection ranging from 40%-60% gave the best results.

Conclusions These results indicate the importance of the choice of a suitable feature selection algorithm and machine learning model.

5. CONCLUSION

This paper offered an extensive investigation into the effect of various feature selection approaches on the machine learning classification accuracy with the utilization of benchmark datasets such as MI, GSE2034,

ISSN 3136-2710

ARCENE, and Changer. The suggested framework was meant to provide a fair and methodical assessment of feature selection techniques by combining data pre-processing, exploratory data analysis, feature selection, machine learning algorithms, cross-validation, and performance evaluation.

It can be stated from the findings that the feature selection procedure is an important step in achieving superior prediction accuracy and other quality metrics. By choosing relevant features only, it is possible to improve the accuracy rate, precision, F1-score, and minimize RMSE values. Hence, it can be concluded that eliminating redundant features positively affects not only prediction quality but also increases efficiency by reducing computational complexity.

In particular, among feature selection strategies applied within this research, Mutual Information and ANOVA demonstrated superior results in all cases. The application of Mutual Information is highly efficient since it takes into account linear and nonlinear relationships between features and classes. In turn, ANOVA showed good performance on those data sets in which there is strong class separability. Chi-Square and Correlation-Based methods contributed to data dimensionality reduction. The analysis of feature subset levels of 20, 40, 60, and 80 indicated that adding additional features does not necessarily contribute to model improvement; on the contrary, moderately sized subsets, namely 40 and 60, proved to be the most effective in terms of both accuracy and speed. Too small feature levels might lead to the loss of essential information, whereas too large – to the inclusion of redundant features.

As far as particular models used in this study, it is necessary to note that SVM and KNN models demonstrated the greatest progress due to feature selection. Specifically, the former proved to be very efficient in working with high dimensional spaces; the latter, in contrast, benefited greatly from decreased noise. Such algorithms as decision tree, Gaussian na¨ıve bayes, and logistic regression demonstrated stable yet mediocre results depending on dataset properties.

The utilization of stratified k-fold cross-validation allowed avoiding biasness due to a consistent check of all models within several data subsets. Moreover, visualization with the help of Tableau facilitated the process significantly. In summary, it is evident that the proper selection of feature selection technique, number of features, and classifier plays a vital role in optimizing the performance of the classifier. It is apparent from the results obtained that feature selection is not just a preprocessing technique but also an important part of the machine learning process.

Future research may involve applying more sophisticated feature selection techniques along with deep learning to achieve better classification performance.

REFERENCES

[1] G. Chandrashekar and F. Sahin, “A survey on feature selection methods,” *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16–28, 2014.

[2] A. Y. Ng, “On feature selection: Learning with exponentially many irrelevant features,” *ICML*, 1998.

[3] I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[4] G. H. John, R. Kohavi, and K. Pfleger, “Irrelevant features and the subset selection problem,” *Machine Learning*, vol. 16, no. 1, pp. 121–147, 1994.

[5] H. Liu and H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*. Springer, 1998.

ISSN 3136-2710

[6] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.
[7] M. A. Hall, *Correlation-based Feature Selection for Machine Learning*. PhD thesis, University of Waikato, 1999.
[8] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinfor-˜ matics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.
[9] D. Dua and C. Graff, "Uci machine learning repository." https://archive.ics. uci.edu/ml, 2019. University of California, Irvine, School of Information and Computer Sciences.
[10] NCBI, "Gene expression omnibus dataset gse2034," 2005.
[11] UCI, "Arcene dataset - uci machine learning repository," 2003.
[12] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *Journal of Machine Learning Research*, vol. 3, pp. 1289–1305, 2003.
[13] K. Fukunaga, "Introduction to statistical pattern recognition," *Academic Press*, 1990.
[14] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2011.
[15] M. Dash and H. Liu, "Feature selection for classification," *Intelligent Data Analysis*, vol. 1, no. 1-4, pp. 131–156, 1997.
[16] P. Langley, "Selection of relevant features in machine learning," *AAAI Fall Symposium*, 1994.
[17] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *Journal of Machine Learning Research*, vol. 5, pp. 1205–1224, 2004.
[18] Y. e. a. Liu, "Chi-square feature selection methods for classification," *ICMLA*, 2010.
[19] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley, 2006.
[20] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria," *IEEE TPAMI*, 2005.
[21] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.
[22] Y. Freund and R. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, 1997.
[23] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273– 297, 1995.
[24] J. Friedman, T. Hastie, and R. Tibshirani, "The elements of statistical learning," 2001.
[25] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation," *IJCAI*, 1995.
[26] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," *IJCAI*, 1995.
[27] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing Management*, vol. 45, no. 4, pp. 427–437, 2009.
[28] D. M. W. Powers, "Evaluation: From precision, recall and f-measure to roc, informedness, markedness and correlation," *Journal of Machine Learning Technologies*, pp. 37–63, 2011.
[29] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (mae) over the root mean square error (rmse)," *Climate Research*, 2005.
[30] C. M. Bishop, "Pattern recognition and machine learning," 2006.
[31] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1-3, pp. 389–422, 2002.