Chuka University
info@chuka.ac.ke | 020 231 0512 / 020 231 0518

SELECTION OF OPTIMAL FEATURES IN STATISTICAL MODELLING

Gachoki, P.K., Njoroge, G.G.[1] and Muraya, M.M.[2]
[1]Department of Physical Sciences, Chuka University, P.O Box 109-60400, Chuka, Kenya [2]Department of Plant Science, Chuka University, P.O Box 109-60400, Chuka, Kenya Corresponding author email: pkgachoki@gmail.com; moses.muraya@chuka.ac.ke

How to cite:
Gachoki, P. K., Njoroge, G. G. and Muraya, M. M. (2021). Selection of optimal features in statistical modelling. *In: Isutsa, D. K. (Ed.). Proceedings of the 7th International Research Conference held in Chuka University from 3rd to 4th December 2020, Chuka, Kenya, p. 555-564*

ABSTRACT
In statistical modelling, selection of optimal features entails making a selection of relevant predictor variables to be used in development of statistical models. Most modelling studies have focused on construction of statistical models skipping out or failing to put on record the process of selection of best features which is an integral part of statistical modeling. This failure might lead to use of duplicated features, features that are less relevant or other that have low variance in addition to random features which could result to poor performing prediction models. This study seeks to discuss how feature selection can be done as a pre-requisite for statistical modeling. Some of the methods used in selection of best features include; forward selection, backward elimination, recursive elimination, entropy selection, variance threshold elimination, chi-square statistics, tree based selection, feature importance and correlation matrix with heat maps. This study is vital to researchers building statistical models since use of optimal features in statistical modeling would lead to high performing statistical models.
Keywords: Feature selection, forward selection, feature importance, correlation matrix with heatmaps

INTRODUCTION

In development of statistical models, features selection entails various objectives. First is obtaining a minimal subset of features that is useful and enough to the concept being targeted (Wang & Liu, 2016). Further it entails selection of a subset of features that can produce an optimum criterion function when compared with all the subsets of the available data. Selection of best features aims at attaining improved prediction accuracy. In addition, feature selection can also aim at getting a reduced structure of the data without a significant decrease in prediction accuracy of the algorithm developed using the optimal selected features (Jovic & Bogunovic, 2015).

Feature selection has continued to be one of the usually skipped steps during development of statistical models (Ramaswami & Bhaskaran, 2009). This process is very vital especially where the data sets involved comprise of many features. The selection of best features can help in dropping out features what are redundant of some that could be highly correlated. The redundancy and correlation could affect the prediction accuracy of the prediction models. This study sought to investigate whether several feature selection methods would help to improve the accuracies of the prediction models (Shardlow, 2016).

Some of the statistical techniques that can be used in feature selection are forward selection, backward elimination, recursive feature elimination, univariate selection, and feature importance and correlation matrix with heatmaps (Venkatesh & Anuradha, 2019). Forward selection is an iterative procedure which starts with having no feature in the model. In each iteration, a feature is added that improves the model till an addition of a new feature does not improve the performance of the model. For backward elimination, the starting point is a model with all the features (Miao & Niu, 2016). In each iteration, the least significant feature is removed until no improvement is observed.

Recursive feature elimination is a greedy otimization algorithm which aims to find the best perfoming feature at each iteration. The algorithm constructs the next model with the features until all the features are exhausted. The features are then ranked based on their order of elimination. However, forward selection, backward elimininationand recursive feature elimination works well with small data sets (Venkatesh & Anuradha, 2019). For big data, the selection process would be

very slow. This necessitates the use of more robust feature selection methods.

Such techniques include; univariate selection, and feature importance and correlation matrix with heatmaps (Duchesnay & Löfstedt, 2018). The general idea underlying these methods is creation of a lot of subsets from the data each time giving the model accuracy. These methods are easy to use and also yield good results. In univariate selection statistical tests can be used to select features that have the strongest relationship with the output variable. This involves selection of the best class that can be used with a suite of different statistical tests to select a specific number of features. For-instance, chisquare statistical tests for non-negative features can be used to select the best 10 features for a certain problem (Duchesnay & Löfstedt, 2018).

Addressing the problem of feature selection in a large data set can also be done using the features importance. This is where the feature importance of each feature in the dataset is obtained using the feature importance property of the model (Helwig, 2017). Feature importance gives a score for each feature of the data. The higher the score the more the importance or the feature is towards the output variable.

Generally, feature importance is an inbuilt class that comes with tree based classifiers. the other method is correlation matrix with heat maps. Correlation states how features are related to each other or the target variable. This correlation is positive if increase in one value of the feature increases the value of the target variable and negative if increase in one value of the feature decreases the value of the target variable (Helwig, 2017).

The objective was achieved by obtaining data with over 32 features and fitting a logistic regression model leaving out feature selection step. The data was then subjected to several feature selection techniques and then refitting the logistic regression model with the reduced data. The feature selection methods applied included; forward selection, backward elimination, recursive feature elimination and entropy feature selection technique. The logistic model was used because the response variable in the data was a binary classification problem. The results showed that features selection produced models with better prediction accuracies when compared with when the process was left out.

METHODOLOGY

The data was obtained the Kaggle website. The response variable was a binary classification of benign and malignant cancer tumors. The data had about 32 explanatory variables. The process of data analysis involved fitting a logistic regression model skipping out the process of feature selection at first.

The prediction accuracy of this model was compared to similar models fitted after subjecting the data to various feature selection techniques. The feature selection methods used were; forward selection, backward elimination, recursive feature elimination and entropy feature selection methods.

Models validation was by using the training and the testing data set. The training data set was 70% of the data and 30% formed the testing sets. Models comparison was done based on the prediction accuracies of the model on the classification problem.

Forward Selection

In the forward selection method, the software looks at all the predictor variables selected and picks the one that predicts the most on the dependent measure (Haque *et al.*, 2018). That variable is added to the model. This is repeated with the variable that then predicts the most on the dependent measure. This little procedure continues until adding predictors does not add anything to the prediction model anymore.

Backward Selection

In the backward selection, all the predictor variables chosen are added into the model. Then, the variables that do not (significantly) predict anything on the dependent measure are removed from the model one by one (Haque *et al.*, 2018). The backward method is the preferred method, because the forward method produces so-called suppressor effects. These suppressor effects occur when predictors are significant when another predictor is held constant.

Recursive Feaure Elimination

This method aims at obtaining a subset of features that yields the best performing model. The method creates models iteratively keeping aside the best and worst performing features after each iteration. Next models are fitted with the remaining features until all the features are exhausted. The ranking of the features is then done based on the order in which they were eliminated.

Entropy Feature Selection

Entropy feature selection is also called mutual information (Glinka *et al*., 2017). This is a basic method that measures how much knowledge between two attributes are correlated. Mathematically, it is defined as;

Consider the high dimensional data D = N*M,
where M is the number of feature and N is the number of the instances.

Let x and y be two random features or variables, p(x) and p(y) be their probability density functions and p(x, y) be their joint probability density function (Singh *et al*., 2014). Then their mutual information (MI) can be defined as; $MI(X,Y) = \sum\sum(,)(,)$

Using the entropy and mutual information the problem can be solved in different ways which are as follows; Let

H(X) denote Shannon's entropy of X, then;

$H(x) = -\int ()\log()()$

The entropy is related to mutual information as follows:
MI(X,Y) = H(X) – H(X,Y)
HI(X,Y) = H(X,Y) – H(X/Y) – H(Y/X)

As a feature selection criterion, the best feature will maximize the mutual information MI(X, Y), where X is the feature vector and Y is the class indicator.

Models Evaluation and Comparison

Models fitted in this study were evaluated on their performance based on the following criteria;

$\text{Accuracy} = \frac{\sum \text{True positive} + \sum \text{True Negative}}{\sum \text{Total Population}}$

$\text{Sensitivity} = \frac{\sum \text{True positive}}{\sum \text{Condition positive}}$

$\text{Specificity} = \frac{\sum \text{True Negative}}{\sum \text{Condition Negative}}$

$\text{Prevalence} = \frac{\sum \text{Condition positive}}{\sum \text{Total Population}}$

$\text{Positive Predicted Value (Precision)} = \frac{\sum \text{True positive}}{\sum \text{Predicted Condition Positive}}$

$\text{Negative Predicted Value} = \frac{\sum \text{True Negative}}{\sum \text{Predicted Condition Negative}}$

$\text{Precision} = \frac{\text{True Positive}}{\text{True positive+false positive}} = \frac{\text{True Positive}}{\text{Total Predicted Positive}}$

$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive+False Negative}} = \frac{\text{True positive}}{\text{Actual Positive}}$

$1\ \text{score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision+Recall}}$

RESULTS AND DISCUSSION
Pleriminary Analysis

The preliminary analysis involved generation of descriptive statistics such as the means, standard deviations, maximums and minimums (Table 1). The results showed that the average radius mean of the tumors was 14.1273 units. The average of the perimeter means of the tumors was 91.969 units. The average of the variable fractal dimension mean was .0627976 units. The mean of the variable smoothness_se was 2.86606 units. The standard deviations of the variables were less than the means. This was a sign of homogeneity among the data values for the various variables. The means for the other variables are as presented in Table 1.

Table 1: Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
	Statistic	Statistic	Statistic	Statistic	Statistic
radius_mean	569	6.98	28.11	14.1273	3.52405
texture_mean	569	9.71	39.28	19.2896	4.30104
perimeter_mean	569	43.8	188.5	91.969	24.2990
area_mean	569	144	2501	654.89	351.914
smoothness_mean	569	.0526	.1634	.096360	.0140641
compactness_mean	569	.0194	.3454	.104341	.0528128
concavity_mean	569	.0000	.4268	.088799	.0797198
concave points_mean	569	.0000	.2012	.048919	.0388028
symmetry_mean	569	.1060	.3040	.181162	.0274143
fractal_dimension_mean	569	.04996	.09744	.0627976	.00706036
radius_se	569	.112	2.873	.40517	.277313
texture_se	569	.3602	4.8850	1.216853	.5516484
perimeter_se	569	.757	21.980	2.86606	2.021855
area_se	569	6.8	542.2	40.337	45.4910
smoothness_se	569	.001713	.031130	.00704098	.003002518
compactness_se	569	.00225	.13540	.0254781	.01790818
concavity_se	569	.00000	.39600	.0318937	.03018606
concave points_se	569	.00000	.05279	.0117961	.00617029
symmetry_se	569	.00788	.07895	.0205423	.00826637
fractal_dimension_se	569	.000895	.029840	.00379490	.002646071
radius_worst	569	7.93	36.04	16.2692	4.83324
texture_worst	569	12.02	49.54	25.6772	6.14626
perimeter_worst	569	50.4	251.2	107.261	33.6025
area_worst	569	185	4254	880.58	569.357
smoothness_worst	569	.0712	.2226	.132369	.0228324
compactness_worst	569	.0273	1.0580	.254265	.1573365
concavity_worst	569	.0000	1.2520	.272188	.2086243
concave points_worst	569	.0000	.2910	.114606	.0657323
symmetry_worst	569	.1565	.6638	.290076	.0618675
fractal_dimension_worst	569	.0550	.2075	.083946	.0180613
Valid N (listwise)	569				

Feature Selection

Feature Selection by Recursive Feature Elimination (RFE)

The process of selecting variables using Recursive Feature Elimination (RFE) is as shown in Table 2. When four best features were selected using RFE, the prediction accuracy of the fitted model was 90.80%. When the best eight features were selected using RFE, the prediction accuracy was 93.97%. After selection of the best sixteen features, the prediction accuracy improved to 94.82%. When thirty best features were selected using RFE, the prediction accuracy of the fitted

model was 94.02%. Therefore, the optimal number of features as selected using RFE was sixteen. The top 5 variables (out of 16) as selected using Recursive Feature Elimination (RFE) were; perimeter_worst, concave_points_worst, area_worst, radius_worst, and concave_points_mean. The list of the selected of the selected features using RFE based on their importance is presented in Table 3.

Table 2: Variable Selection by Recursive Feature Elimination (RFE)

Variables	Accuracy	Kappa	AccuracySD	KappaSD	Selected
4	0.9080	0.8044	0.03563	0.07491	
8	0.9397	0.8715	0.03133	0.06639	
16	0.9482	0.8900	0.03104	0.06576	*
30	0.9402	0.8725	0.03295	0.07047	

Table 3: Features selected using Recursive Feature Elimination (RFE) based on their importance

[1]"perimeter_worst"	[2]"concave_points_worst"	[3]"area_worst"
[4]"radius_worst"	[5] "concave_points_mean"	[6]"area_se"
[7]"texture_worst"	[8] "concavity_worst"	[9]"texture_mean"
[10]"concavity_mean"	[11]"area_mean"	[12]"radius_se"
[13]"smoothness_worst"	[14]"perimeter_mean"	[15]"perimeter_se"
[16]"radius_mean"		

Selection of Features by Backward Selection

The process of selection of features by backward selection was similar to the RFE. From backward selection, the optimal number of features was eighteen (Table 4). This was two features more than what had been selected using RFE. The list of the features selected using backward selection is presented in Table 5.

Table 4: Number of features selected by backward selection

Samples	Predictors selected	Classes
398	18	2

Table 5: Features Selected by Backward Selection based on their importance

	DF	Deviance	AIC
<none>		0.00	38.00
perimeter_mean	1	22.05	58.05
compactness_worst	1	23.08	59.08
concavity_mean	1	25.78	61.78
radius_mean	1	26.20	62.20
concavity_se	1	28.18	64.18
symmetry_mean	1	28.24	64.24
compactness_se	1	33.27	69.27
concave_points_se	1	34.41	70.41
fractal_dimension_worst	1	34.77	70.77
symmetry_se	1	36.59	72.59
concave_points_mean	1	37.00	73.00
fractal_dimension_se	1	38.01	74.01
symmetry_worst	1	38.94	74.94
radius_worst	1	39.52	75.52
compactness_mean	1	41.26	77.26
area_se	1	42.74	78.74
texture_mean	1	44.61	80.61
perimeter_worst	1	1081.31	1117.31

Selection of Features by Entropy Selection

When the entropy selection was used, twenty five features were selected from the data (Table 6). The selected features were listed based on their importance in Table 7. The features were then used to fit models so that comparison can be with

models fitted using features selected using other methods.

Table 6: Number of features selection by entropy selection

Samples	Predictors selected	Classes
398	25	2

Model Fitting

Logistic Regression Model with all the Features

The results in Table 9 is the logistic regression model where feature selection had not been done. The model had all the 32 features from the data. The results indicated that an increase in the variable radius mean of the tumor by 1 unit decreased the chances of the tumor being malignant by 9.903775e+02 times. A unit increase in the texture mean by 1 unit increased the chances of the tumor being malignant by 1.145513e+01 times. A unit increases in the concave points mean by 1 unit increases the chances of the tumor being malignant by 1.138650e+04 times. A unit increase in the fractal dimension mean by 1 unit decreases the chances of a tumor being malignant increases by 3.493666e+03 times. A unit increases in the variable perimeter_se increases the chances of a tumor being malignant by 5.366880e+01 times. Variables investigated had a significant effect on the type of tumor being diagonised.

Selection of Features by Forward Selection

Use of forward selection to select optimal features yielded the highest number of features. The number of features selected was thirty as presented in Table 8. The results in Table 10 shows how the fitted logistic regression model performed in classifying and predicting if a tumor was malignant or benign. The model had a prediction accuracy of 97.08%. The precision (positive predicted value) of the model was 98.25%. The model sensitivity was 93.33%.

Table 7: List of features by their importance using entropy feature selection

Attributes	**Importance**
perimeter_worst	0.4850561
area_worst	0.4675581
concave_points_worst	0.4538449
radius_worst	0.4478213
concave_points_mean	0.4155797
perimeter_mean	0.4087355
area_mean	0.3881128
radius_mean	0.3814810
area_se	0.3664849
concavity_mean	0.3499271
concavity_worst	0.3458024
radius_se	0.2562297
perimeter_se	0.2523637
compactness_worst	0.2145325
compactness_mean	0.2142234
concavity_se	0.1483622
concave_points_se	0.1402913
texture_mean	0.1265121
texture_worst	0.1217746
symmetry_worst	0.1008219
smoothness_worst	0.0941130
compactness_se	0.0691604
symmetry_mean	0.0669995
smoothness_mean	0.0641805
fractal_dimension_worst	0.0596582
symmetry_se	0.0272433
fractal_dimension_se	0.0257642
fractal_dimension_mean	0.0231045
texture_se	0.0000000
smoothness_se	0.0000000

Table 8: Number of features selection by forward selection

Samples	Predictors selected	Classes
398	30	2

Model Fitting using Variables from Forward Selection

The performance of the model fitted using forward selection was presented in Table 11. The model performance based on accuracy, precision and sensitivity was 97.08%, 98.25% and 93.33%. This performance was similar to the model fitted without doing any feature selection. However, this model had 30 features as compared to the model without feature selection that had 32 features. The reduction in the number of features led to reduced training time of the model. This results indicated that it was possible to drop some features that could have been redundant or highly correlated with the remaining features without significantly lowering the model performance.

Model Fitting using Variables from Entropy Selection

The variables selected using entropy selection were applied in model fitting to find out if the performance of the model was improved. The results were presented in Table 12. The model performance was 95.91% in terms of accuracy, 92.06% in terms in terms of precision and 96.67% in terms of sensitivity. This was a reduced performance when compared to the model without feature selection and the model from the forward selected variable. Entropy selection yielded twenty five features which was lower that the complete set of features or the features from forward selection. A reduction of the prediction accuracy meant that this method dropped some features that were vital in model building.

Table 9: Logistic Regression Model

Term	Estimate	std.error	Statistic
(Intercept)	1.051554e+02	4.009638e+05	0.0002623
radius_mean	-9.903775e+02	1.560410e+05	-0.0063469
texture_mean	1.145513e+01	3.530283e+03	0.0032448
perimeter_mean	9.743985e+01	2.559560e+04	0.0038069
area_mean	2.585920e+00	5.745781e+02	0.0045006
smoothness_mean	2.947579e+03	1.035082e+06	0.0028477
compactness_mean	-8.526539e+03	8.904460e+05	-0.0095756
concavity_mean	2.219474e+03	4.943034e+05	0.0044901
concave_points_mean	1.138650e+04	9.282659e+05	0.0122664
symmetry_mean	-2.836263e+03	2.343713e+05	-0.0121016
fractal_dimension_mean	-3.493666e+03	1.652806e+06	-0.0021138
radius_se	-1.635119e+03	5.034541e+05	-0.0032478
texture_se	-1.992183e+01	2.387207e+04	-0.0008345
perimeter_se	5.366880e+01	2.473653e+04	0.0021696
area_se	2.032495e+01	3.870656e+03	0.0052510
smoothness_se	-2.378366e+04	2.990972e+06	-0.0079518
compactness_se	1.631593e+04	3.105827e+06	0.0052533
concavity_se	-6.128921e+03	5.893287e+05	-0.0103998
concave_points_se	3.931166e+04	2.950471e+06	0.0133239
symmetry_se	-2.073166e+04	2.814956e+06	-0.0073648
fractal_dimension_se	-1.088166e+05	1.888605e+07	-0.0057617
radius_worst	3.877885e+02	4.104385e+04	0.0094482
texture_worst	1.384641e+00	3.685304e+03	0.0003757
perimeter_worst	-2.947268e+01	5.744130e+03	-0.0051309
area_worst	-1.043073e+00	3.184540e+02	-0.0032754
smoothness_worst	-1.177138e+03	4.179253e+05	-0.0028166
compactness_worst	-1.178895e+03	2.761519e+05	-0.0042690
concavity_worst	3.930469e+02	1.677193e+05	0.0023435
concave_points_worst	-1.419982e+03	5.219040e+05	-0.0027208
symmetry_worst	3.534347e+03	2.750049e+05	0.0128519
fractal_dimension_worst	1.190558e+04	1.485167e+06	0.0080163

Table 10: Performance of the model fitted without feature selection

Accuracy	0.9708
95% CI	(0.9331, 0.9904)
No Information Rate	0.6491
P-Value [Acc > NIR]	<2e-16
Kappa	0.9351
Mcnemar's Test P-Value	0.3711
Sensitivity	0.9333
Specificity	0.9910
Pos Pred Value	0.9825
Neg Pred Value	0.9649
Prevalence	0.3509
Detection Rate	0.3275
Detection Prevalence	0.3333
Balanced Accuracy	0.9622

Model Fitting using Features from Recursive Feature Selection

The features selected using Recursive Feature Selection were used in model fitting and performance compared with models fitted using features from other methods. The model had 95.91% accuracy, 85.55% precision and a sensitivity of 100%. This performance was similar to the model fitted with features from entropy selection. However, recursive feature selection yielded fewer features which made the model to train faster compared to all the other models. The model fitted using features from RFE performed lower than the model from no feature selection, backward elimination and forward selection.

Table 11: Performance of the model fitted using variables from forward selection

Accuracy	0.9708
95% CI	(0.9331, 0.9904)
No Information Rate	0.6491
P-Value [Acc > NIR]	<2e-16
Kappa	0.9351
Mcnemar's Test P-Value	0.3711
Sensitivity	0.9333
Specificity	0.9910
Pos Pred Value	0.9825
Neg Pred Value	0.9649
Prevalence	0.3509
Detection Rate	0.3275
Detection Prevalence	0.3333
Balanced Accuracy	0.9622

Table 12: Perfomance of the model fitted using variables from entropy selection

Accuracy	0.9591
95% CI	(0.9175, 0.9834)
No Information Rate	0.6491
P-Value [Acc > NIR]	<2e-16
Kappa	0.9112
Mcnemar's Test P-Value	0.4497
Sensitivity	0.9667
Specificity	0.9550
Pos Pred Value	0.9206
Neg Pred Value	0.9815
Prevalence	0.3509
Detection Rate	0.3392
Detection Prevalence	0.3684
Balanced Accuracy	0.9608

Table 13: Perfomance of the model fitted using variables from recursive feature selection

Accuracy	0.9591
95% CI	(0.9175, 0.9834)
No Information Rate	0.6491
P-Value [Acc > NIR]	<2e-16
Kappa	0.9125
Mcnemar's Test P-Value	0.2334
Sensitivity	1.0000
Specificity	0.9369
Pos Pred Value	0.8555
Neg Pred Value	1.0000
Prevalence	0.3509
Detection Rate	0.3509
Detection Prevalence	0.3918
Balanced Accuracy	0.9685

Model Fitting using Features from Backward Selection

The results of the performance of the model fitted with features selected by backward elimination is presented in Table 14. The number of features was eighteen. The model fitted had performance similar to the model fitted without features selection and the model fitted using forward selection. However, the models fitted without feature selection and the one for forward selection took more time to train due to the more number of features involved. This means that the model by from backward eliminated features was better compared to the other models. Of all the models fitted, this is the model that attained the best prediction accuracy taking less time to train.

Table 14: Performance of the model fitted using variables from backward elimination

Accuracy	0.9708
95% CI	(0.9331, 0.9904)
No Information Rate	0.6491
P-Value [Acc > NIR]	<2e-16
Kappa	0.9351
Mcnemar's Test P-Value	0.3711
Sensitivity	0.9333
Specificity	0.9910
Pos Pred Value	0.9825
Neg Pred Value	0.9649
Prevalence	0.3509
Detection Rate	0.3275
Detection Prevalence	0.3333
Balanced Accuracy	0.9622

Models Comparison

The models fitted using data from various methods of feature selection were compared using the F-score values which incorporated both the precisions of the models in addition to their recall values. The results were presented in Table 15. The results indicated that backward selection yielded features that produced the best performing model in terms of accuracy, F1 score and training speed. Models from no feature selection and forward selection has similar values of accuracy and F1 scores but took more time to train due to the increased number of features.

Recursive feature selection yielded the least number of features after selection and thus training a model using the features took the least time. However, the model produced less accuracy F1 scores as compared to the model from backward eliminated features. This meant that the method dropped some features which were key in improving the model performance.

Table 15: Comparing Performance of the Fitted Models

Model	No. of Features	Accuracy	Precision (positive predicted value)	Recall (sensitivity)	F1 Score
No feature selection	32	0.9708	0.9825	0.9333	0.957268
Forward selection	30	0.9708	0.9825	0.9333	0.957268
Backward selection	18	0.9708	0.9825	0.9333	0.957268
Entropy selection	25	0.9591	0.9206	0.9667	0.943087
Recursive feature selection	16	0.9591	0.8555	1.0000	0.922123

CONCLUSION

In conclusion, the aim of this study was to review some features selection methods and find out whether feature selection is an important prerequisite in statistical modeling. The methods tackled were forward selection, backward elimination, entropy selection and recursive feature selection method. The findings revealed that feature selection was important since it yield some fewer features which could still relay the same information contained in the complete set features. This is because the dropped features were either highly correlated to the retained set of features or they were redundant. Such features could therefore be dropped without losing the information from the complete set of features. Features selection also helped to reduce the model training time. From the methods applied in in this study, backward selection produced the best performance in selecting the best features. The method reduced the number of features from thirty two to eighteen and still retained the prediction accuracy of the model with all the features. Backward elimination therefore outperformed the other considered methods when applied with the data set considered for this study.

REFERENCES

Duchesnay, E., & Löfstedt, T. (2018). Statistics and Machine Learning in Python. *Release 0.1*.

Helwig, N. E. (2017). Data, Covariance, and Correlation Matrix. *University of Minnesota (Twin Cities)*.

Jović, A., Brkić, K., & Bogunović, N. (2015, May). A review of feature selection methods with applications. In *2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO)* (pp. 1200-1205). Ieee.

Miao, J., & Niu, L. (2016). A survey on feature selection. *Procedia Computer Science*, *91*, 919-926.

Ramaswami, M., & Bhaskaran, R. (2009). A study on feature selection techniques in educational data mining. arXiv preprint arXiv:0912.3924.

Shardlow, M. (2016). An analysis of feature selection techniques. *The University of Manchester*, *1*, 1-7.

Venkatesh, B., & Anuradha, J. (2019). A review of feature selection and its methods. *Cybernetics and Information Technologies*, *19*(1), 3-26.

Wang, Suhang & Tang, Jiliang & Liu, Huan. (2016). Feature Selection. 10.1007/978-1-4899-7502-7_101-1.

Haque, M. M., Rahman, A., Hagare, D., & Chowdhury, R. K. (2018). A comparative assessment of variable selection methods in urban water demand forecasting. *Water*, *10*(4), 419.

Glinka, K., Woźniak, R., & Zakrzewska, D. (2017, June). Improving multi-label medical text classification by feature selection. In *2017 IEEE 26th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)* (pp. 176-181). IEEE.

Singh, B., Kushwaha, N., & Vyas, O. P. (2014). A feature subset selection technique for high dimensional data using symmetric uncertainty. *Journal of Data Analysis and Information Processing*, *2*(04), 95.

****