Knowledge and Information Systems (2024) 66:1955–1995
https://doi.org/10.1007/s10115-023-01993-5

REGULAR PAPER

Supervised feature selection using principal component analysis

Fariq Rahmat[1] · Zed Zulkafli[2] · Asnor Juraiza Ishak[1] · Ribhan Zafira Abdul Rahman[1] · Simon De Stercke[3] · Wouter Buytaert[3] · Wardah Tahir[4] · Jamalludin Ab Rahman[5] · Salwa Ibrahim[6] · Muhamad Ismail[6]

Received: 28 June 2022 / Revised: 26 July 2023 / Accepted: 15 September 2023 /
Published online: 8 November 2023

Abstract

The principal component analysis (PCA) is widely used in computational science branches such as computer science, pattern recognition, and machine learning, as it can effectively reduce the dimensionality of high-dimensional data. In particular, it is a popular transformation method used for feature extraction. In this study, we explore PCA's ability for feature selection in regression applications. We introduce a new approach using PCA, called Targeted PCA to analyze a multivariate dataset that includes the dependent variable—it identifies the principal component with a high representation of the dependent variable and then examines the selected principal component to capture and rank the contribution of the non-dependent variables. The study also compares the feature selected with that resulting from a Least Absolute Shrinkage and Selection Operator (LASSO) regression. Finally, the selected features were tested in two regression models: multiple linear regression (MLR) and artificial neural network (ANN). The results are presented for three socioeconomic, environmental, and computer image processing datasets. Our study found that 2 of 3 random datasets have more than 50% similarity in the selected features by the PCA and LASSO regression methods. In the regression predictions, our PCA-selected features resulted in little difference compared to the LASSO regression-selected features in terms of the MLR prediction accuracy. However, the ANN regression demonstrated a faster convergence and a higher reduction of error.

Keywords Supervised feature selection · Feature selection · LASSO · Principal component analysis · ANN

Fariq Rahmat, Zed Zulkafli, Asnor Juraiza Ishak, Ribhan Zafira Abdul Rahman, Simon De Stercke, Wouter Buytaert, Wardah Tahir, Jamalludin Ab Rahman, Salwa Ibrahim, and Muhamad Ismail contributed equally to this work.

✉ Zed Zulkafli
zeddiyana@upm.edu.my

Extended author information available on the last page of the article

1 Introduction

Feature selection is widely used in computational science branches, such as computer science, pattern recognition, and machine learning, to effectively reduce high-dimensional data. Feature selection can improve, firstly, computational efficiency and, secondly, the accuracy of the prediction algorithms [1]. Three major traditional feature selection approaches for machine learning development include the filter, wrapper, and embedded methods [2]. The filter method selects features based on certain evaluation criteria, such as a high joint probability or correlation between input and output variables [3–8]. Meanwhile, the wrapper method conducts feature selection through the machine learning algorithm, which evaluates all possible combinations of features by using a searching strategy and produces the result in a machine learning of the training dataset [9]. Lastly, the embedded method is similar to the wrapper method but derives the features during model training via a regulation technique that adds a penalty to the different parameters of a model to reduce its freedom [10].

Feature selection differs from feature extraction in that the former creates a subset of the initial inputs, while the latter produces new composite features. Feature extraction is at times undesirable as its transformation of initial features removes their identifiability. A powerful and commonly used method for feature extraction method is the principal component analysis (PCA). By contrast, few published works exist on the implementation of PCA for feature selection. One previous study investigated the contribution of features toward the principal components (PC) with the largest eigenvalues [11]. This contribution value is the relative measure of a feature's representation quality for the selected PC over the total representation quality of all features. The features were sorted in descending order of contribution, and their ranks were considered an indicator of relative importance [11]. Another study in 2018 applied a similar method for feature selection but only selected the first two highest correlation coefficients from each selected PC [12]. In the same year, a group of researchers from China applied a new method to implement PCA for feature selection on high-dimensional data before they could be applied to the clustering model [13]. The method first reduces the dimensionality of the data using a robust PCA technique that is less sensitive to outliers than traditional PCA. Robust PCA is a dimensionality reduction technique that aims to extract the most important features while minimizing the influence of outliers. This is achieved by decomposing the data matrix into low-rank and sparse components, where the low-rank component captures the underlying structure of the data and the sparse component accounts for the outliers. This way, the method automatically identifies and selects the most important features while minimizing the impact of noisy or irrelevant features [13]. Once the dimensionality of the data is reduced, the local adaptive learning algorithm is applied to learn the clustering structure of the reduced-dimensional data. The adaptive learning algorithm adaptively adjusts the bandwidth of the kernel function used for density estimation, allowing it to capture the local structure of the data. All three studies involve unsupervised feature selection for pattern recognition and image processing applications.

Our study aimed to adapt these approaches to the supervised feature selection problem. We introduce a new approach using PCA, called Targeted PCA to analyze a multivariate dataset that includes the dependent variable. The reviewed studies [11–13] determined the selection of the PC based on explained variance and the rank of contribution along the selected PC governed feature selection in unsupervised learning applications. Guided by this, we explored the implementation of the same method but also considered the dependent variable within the dataset for supervised learning applications. The method can be summarized in three parts. Firstly, it performs PC selection based on variance explained exceeding a certain threshold. Secondly, it selects one or more reference PC(s) based on a top contribution rank by the dependent variable. Lastly, it finalizes feature selection based on contribution values exceeding a certain threshold from among the independent variables on the reference PCs. The approach is assessed in two ways: First, the selected features are compared with features selected using the LASSO regression model. Second, they were used as input in linear (multiple linear regression) and nonlinear (artificial neural network) regression models. We used three datasets covering socioeconomic, environmental, and computer image processing fields of applications.

2 Materials and methods

The full methodology of Targeted PCA is presented in Fig. 1, and detailed descriptions are presented in the following subsections. The final section (ref Dataset section) describes three datasets that were used to evaluate the methodology.

2.1 Method development

This section presents the proposed modification to PCA for feature selection. The process begins with a standard calculation of eigenvalues λ and eigenvectors v based on the covariance matrix W as represented by Eq. 1.

$$Wv = \lambda v \tag{1}$$

The eigenvalues λ and eigenvectors v can be solved by rearranging eq. 1 into eq. 2, where I is the identity matrix, then applying the singular value decomposition (SVD) technique.

$$(W - \lambda I)v = 0 \tag{2}$$

The following steps are used to perform the feature selection:

1. Identify and select the PCs (i.e., the eigenvectors) with individual variance explained percentage higher than 1% and cumulative variance explained percentage at minimum 80%. According to Hair (2009), PCA has no universal minimal cumulative explained variance [14]. Instead, the explained variance is based on the analysis context and desired level. Therefore, we chose 80% as the threshold for cumulative explained variance, a common percentage value in many previous studies [11, 12]. Meanwhile, we chose 1% for the threshold of variance explained based on a previous study by Mubarak et al. (2018). The previous study also suggested not selecting too low a threshold because it may include many PCs and increase the complexity of the feature selection process.
2. Identify the quality of representation, $R^2_{j,p}$, of feature components toward the PC [15]. Since all features are represented in the form of a geometrical coordinate, this is determined from the cosine rule, which dictates that for any given variable vector, $R^2_{j,p}$ is equal to the squared cosine of the angle θ between the vector of a selected principal component and given variable vector. A higher $R^2_{j,p}$ value indicates a smaller θ, hence a good representation of the variable on the principal component. This is illustrated in Fig. 2. The main reasons for the selection of the squared cosine in principal component analysis (PCA) in measuring the quality of representation of data features are as stated below:
    (a) It measures the angle between variable and PC vectors, rather than their magnitude, making it robust to scale differences. This is important in PCA because the magnitude

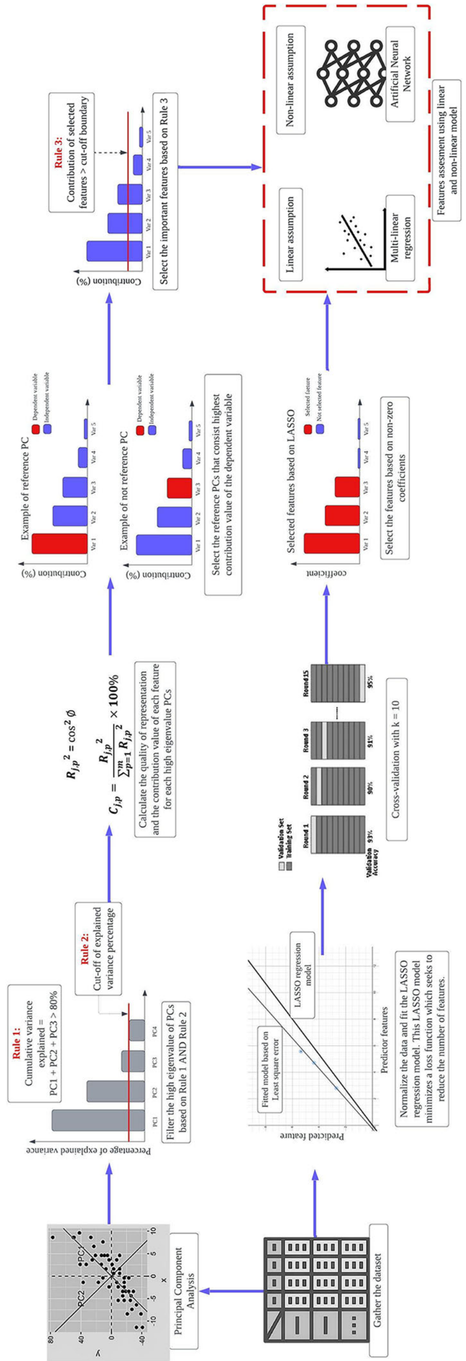

Fig. 1 The illustration of the total framework of Targeted PCA and the LASSO regression model

of the vectors in the PCA space may differ from those in the original space due to the transformation process.

(b) It is robust to outliers. According to Abdi and Williams (2010), the squared cosine similarity metric is robust to outliers because it penalizes large angles between vectors more heavily than small angles [16]. For example, consider two vectors, v1 and v2, with an angle of 60 degrees between them. The cosine similarity value between these vectors is 0.5. However, the squared cosine similarity value is 0.25, smaller than the original cosine similarity value. If the angle between the vectors is smaller, e.g., 30 degrees, the cosine similarity value would be 0.87, and the squared cosine similarity value would be 0.76, closer to 1.

(c) Other methods may be used to measure the similarities between these two parameters, such as Pearson correlation, Euclidean distance, Manhattan distance, and Mahalanobis distance. However, cosine squared has a straightforward interpretation and is easy to compute. According to Kassambara (2017), the cosine rule is the most common practice used in calculating the quality of representation of variables in each PC.

3. Identify the contribution value of each feature to each selected PC from the relative quality of representation (Eq. 3, where $j = 1, 2, \ldots,$ total number of PC and $p = 1, 2, \ldots, m$. m is the total number of features in the dataset).

$$C_{j,p} = \frac{R_{j,p}^2}{\sum_{p=1}^{m} R_{j,p}^2} \times 100\% \tag{3}$$

4. Select the PC corresponding to the largest $C_{j,p}$ of the dependent variable data as the reference PC.
5. Calculate a cutoff point for the relative contribution value as shown in Eq. 4 [following 15]. The cutoff parameter can be calculated as an expected (average) contribution. If the variables' contribution were equal, the expected value would be divided by the total number of variables, m.

$$C_{\text{off}} = \frac{1}{m} \times 100\% \tag{4}$$

6. Select the features with the contribution value, $C_{j,p}$ higher than cutoff value, C_{off}, contributing to the reference PC.
7. Rank the importance of each feature toward the reference PC by comparing the $C_{j,p}$ value, as obtained in Step 3. Ranking the features according to the contribution value in descending order may expedite the filtering process using the threshold method (explained in the previous subsection under point number 6). The higher the $C_{j,p}$, the higher the correlation between the feature to the PC and, thus, the dependent variable. However, a limitation comes when more than one reference PC is selected. The rank for all features cannot be determined based on the $C_{j,p}$ across all reference PCs because they carry different information. Thus, the features are ranked separately for each reference PC.

2.2 Rationale

The Targeted PCA is the new method in feature selection, an evolution of traditional PCA. This section explains the justification for the proposed method based on the original principles of PCA and demonstrates the advantages of Targeted PCA in the feature selection procedure. An established principle of the PCA is that the eigenvector corresponding to a larger eigenvalue can capture more representative sample information [17]. For this reason, it

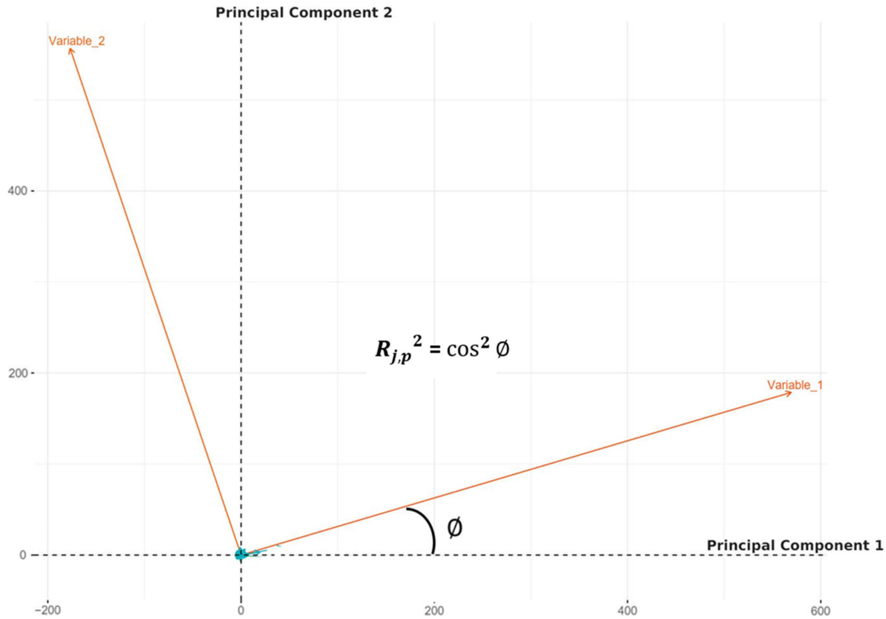

Fig. 2 Implementation of cosine rule in the calculation of quality of representation

is reasonable to investigate the eigenvectors corresponding to larger eigenvalues when one is interested in explaining the variance of the data along each feature's axis. Analyzing multiple eigenvectors allows for a more robust evaluation, considering multiple angles and directions of dependencies. Our proposed method considers analyzing more than one PC, but only those with significant $C_{j,p}$ of the dependent variable. We leverage this property to improve the filtering of the features without losing the information on the correlation between dependent and independent variables. Next, we assess the $C_{j,p}$ of each feature component in the PC that can explain its importance and relation toward the reference PC [11, 12]. The computation of this value accounts for the importance and relation of all features toward the same reference PC. By extension, their importance and relation toward each other are accounted.

2.3 Evaluation

Validating a new method with established methods allows an objective evaluation of its performance. Furthermore, it allows an analysis into the strengths and weaknesses of the different methods compared, facilitating the identification of gaps and opportunities for future research. Assessment is conducted by (1) analysis of the features selection by the Targeted PCA with that of an established feature selection method, the Least Absolute Shrinkage and Selection Operator (LASSO) regression, and (2) measuring the ability of selected features to fit linear and nonlinear models.

2.3.1 Analysis of Selected Features

The Least Absolute Shrinkage and Selection Operator (LASSO) regression

LASSO was introduced by Tibshirani [18]. The regression method minimizes the least squares and has an additional penalty/regularization term for the regression coefficients based on the L1-Norm. The LASSO estimate is defined by the solution to the L1 optimization problem, which is to minimize $\left(\frac{\|Y-X\beta\|_2^2}{n}\right)$, subject to $\sum_{j=1}^{k}\|\beta\|_1 < t$, where t is the upper bound for the sum of coefficients in Eq. 5. Suppose X and Y are the input and output vectors, respectively, β is the vector of the coefficients for all features, k is the number of features, and n is the total number of samples.

$$\hat{\beta}(\lambda) = \underset{\beta}{\operatorname{argmin}}\left(\frac{\|Y - X\beta\|_2^2}{n} + \lambda\|\beta\|_1\right) \tag{5}$$

where $\|Y - X\beta\|_2^2 = \sum_{i=0}^{n}\left(Y_i - (X\beta)_i\right)^2$, $\|\beta\|_1 = \sum_{j=1}^{k}\|\beta\|_1$ and $\lambda > 0$ is the parameter that controls the strength of the penalty—the larger the value of λ, the greater the amount of shrinkage.

The relationship between λ and the upper bound t is an inverse one. As t tends toward infinity, the problem becomes an ordinary least square, and λ becomes 0. Conversely, as t tends toward 0, all coefficients reduce toward 0, while λ goes to infinity. This yields LASSO its variable selection capability—as we minimize the error in the optimization algorithm, some coefficients are shrunk to zero, i.e., $\hat{\beta}_j(\lambda) = 0$, for some values of j (depending on the value of the parameter λ). In this way, the features with coefficients equal to zero are excluded from the model.

The cross-validation (CV) for standard LASSO utilizes the cv.glmnet implementation in R that provides efficient minimization by path-wise coordinate descent for coefficient updates and a method called 'covariance update,' which is a dynamic programming approach to increase the efficiency of the solver [18].

The necessary parameters are:

- *nfolds* = 10 is the number of folds used for the CV.
- *keep* = TRUE makes sure that the information about the fold selection is stored. Since the folds are generated randomly, this was a necessary adjustment.
- *family* = 'Gaussian' is the option for ordinary regression for linear labels.
- *type.measure* = 'mse' (mean squared error) is the indicator for the evaluation method. It measures the deviation from the fitted mean to the response.
- *alpha* = 1 is a hyperparameter that denotes the elastic-net mixing that the study could use if a L1 and L2 penalty mixture is wanted. alpha = 0 is used for ridge regression(L2) and alpha = 1 for pure LASSO regression. The increasing number of alpha may reduce the number of selected features.
- A fitted LASSO model is used to compute the best coefficient value for each independent variable.

Comparability of Targeted PCA and LASSO regression

A LASSO regression is conducted for validating the PCA as both are similar in their function and approach. Firstly, both PCA and LASSO regression can effectively reduce the dimensionality of the feature space. They aim to filter and select a subset of features that capture the most relevant information for predicting the target variable while discarding less important or redundant features. Secondly, both techniques implicitly rank the features based on their importance. In PCA, the principal components are ranked in descending order of the explained variance they capture. Features with high loadings in the top-ranked components are considered more influential. In LASSO regression, the features with nonzero coefficients are deemed important for prediction, while those with zero coefficients are considered less relevant. Lastly, PCA and LASSO regression both operate on linear combinations of features. PCA creates linear combinations (principal components) of the original features, while LASSO regression finds the optimal linear combination of the features as predictors.

Comparison of selected features

The selected features by Targeted PCA and LASSO regression are compared in terms of (1) the number of selected features and (2) the similarities and differences of selected and non-selected features.

To measure the similarities of selected features, we used the Hamming distance technique [19]. This technique is often used to quantify the extent to which two-bit strings of the same dimension differ. In a traditional application of the Hamming distance, the only concern is whether the corresponding bits in two strings agree. However, over the past few years, many researchers have started implementing this method in data preprocessing for machine learning [20, 21]. The Hamming distance is used to find the pairwise similarity in the input space to avoid the excessive redundancies of the input sample.

In this case study, we generalize all the features into bit strings depending on the total number of features used in the dataset:

1. We create two-bit strings representing all selected features from the suggested PCA and LASSO regression methods.
2. We measure the similarity of bits from both bit strings.
3. We calculate the similarity percentage by dividing the total number of similar bits by the length of bit strings.

2.3.2 Linear and nonlinear modeling

Next, two learning algorithms were fitted using selected features from the Targeted PCA and LASSO regression, and their modeling performance was comparatively assessed to establish any advantage of the Targeted PCA. Both learning algorithms are briefly described in the following subsections.

Multiple Linear Regression

In multiple linear regression analysis, an attempt is made to account for the variation of the independent variables with respect to the dependent variable synchronously [22]. The regression analysis model is formulated as in Eq. 6.

$$y = X_1\beta_1 + X_2\beta_2 + \ldots + X_k\beta_k + \epsilon \tag{6}$$

where y denotes the dependent (or study) variable that is linearly related to k independent (or explanatory) variables $X_1, X_2, \ldots, X_k$ through parameters $\beta_1, \beta_2, \ldots, \beta_k$. The parameters $\beta_1, \beta_2, \ldots, \beta_k$ are the regression coefficients associated with $X_1, X_2, \ldots, X_k$, respectively, and ϵ is the random error component reflecting the difference between the observed and fitted linear relationship. There can be various reasons for such differences, e.g., the joint effect of those variables not included in the model, random factors that cannot be accounted for, etc. In a regression equation, the ϵ random error refers to the residual variation that the model does not explain. Furthermore, the ϵ parameter has also been used in LASSO regression to include the bias characteristic in the fitted model.

Metrics R^2 and adjusted R^2 have been used in this study. R^2 measures the proportion of variance in the dependent variable explained by the regression model. It ranges from 0 to 1, with higher values indicating a better fit. R^2 is calculated as the ratio of the sum of squared

Table 1 Standard setup for the experimental setting of ANN model for train and test selected dataset

Elements	Experimental setting
Input neuron	Selected features using PCA or LASSO regression
Number of hidden layers	1
Number of hidden neurons	Input number × 2
Activation function	Sigmoid function(hidden) & Linear function(output)
Optimization algorithm	Stochastic Gradient descent
Learning rate	0.01
Stopping rule 1	Early stopping algorithm (threshold = 0.001)
Stopping rule 2	1,000,000 iterations
Error function	Sum of squared error

errors (SSE) of the regression model to the total sum of squares (SST) of the data:

$$R^2 = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}} \tag{7}$$

where SSE is the sum of squared errors between the predicted and observed values of the dependent variable, and SST is the total sum of squares of the dependent variable.

Adjusted R^2, on the other hand, takes into account the number of predictor variables in the model. It adjusts R^2 by penalizing the addition of extra predictor variables that do not significantly improve the fit of the model. Adjusted R^2 is calculated as:

$$\text{Adjusted } R^2 = 1 - \left[\frac{(1 - R^2) \times (n - 1)}{n - p - 1}\right] \tag{8}$$

where n is the sample size, and p is the number of predictor variables in the model.

Artificial Neural Network

ANN is composed of elementary computational units called neurons combined according to different architectures with multiple numbers of layers of network [23]. They are also known as generalized nonlinear models. Typically, the model performance of the ANN changes depending on model hyperparameter tuning and training dataset manipulation [23]. Thus, to analyze the impact of selected features, the experimental settings were set constant to avoid that additional bias is introduced affecting the model performance.

Table 1 presents the experimental setting of the ANN model used to evaluate the regression of the dependent output data on the selected features. Table 1 presents the experimental setting of the ANN model used to evaluate the regression of the dependent output data on the selected features. Two stopping rules were used for the ANN model training. The first rule applied the early stopping algorithm, which monitored loss in mean squared error (MSE) over time (epochs), and stopped the training when the difference in the loss between previous and current epochs was lower than a threshold value set at 0.001, and the loss increased again in the following epoch. The second rule avoids that the number of epochs keeps increasing due to a non-converging model by stopping the training if the iteration numbers reach 1,000,000.

The ANN can capture the nonlinearity in the dataset because of the activation function used in the algorithm. We use 70%, 10%, and 20% of the dataset for training, validation, and testing stages, respectively. Thus, the ANN can maintain the generalization of patterns in the dataset while also identifying the nonlinearity connection between input and output variables [24].

Table 2 Description of the case dataset

Data characteristic	Description		
Dataset notation	Dataset 1	Dataset 2	Dataset 3
Name	Communities and crime dataset	Relative location of CT slices on axial axis dataset	Leptospirosis incidence and land-use types dataset
Dataset characteristic	Multivariate	Multivariate	Multivariate
Attribute numbers	100	385	215
Associated task	Regression	Regression	Regression
Sample numbers	1994	53,500	513
Missing value	No	No	No
Area	Socioeconomic	Computer image processing	Epidemiology and environment

2.4 Dataset

The study used two public-domain datasets from the UCI Machine Learning Repository collection. A third dataset was from the Federal Department of Town and Country Planning Peninsular Malaysia and the Ministry of Health Malaysia.

The first dataset combines socioeconomic data from the 1990 US Census, law enforcement data from the 1990 US LEMAS survey, and crime data from the 1995 FBI UCR [25]. The second dataset is the medical dataset retrieved from 53,500 computed tomography (CT) images from 74 patients (43 male, 31 female). This dataset predicts the CT slice's relative location on the human body's axial axis [26]. These data are represented in histogram analysis of CT values which describe the bone structures (from value0 to value239) and air inclusion (from value240 to value383). The third dataset is an environmental dataset consisting of 215 land-use types that predict the number of leptospirosis cases that occur in Negeri Sembilan, Malaysia. Land-use types include agriculture, jungle, sport and recreational areas, public infrastructure, and residential areas. Each sample in this dataset represents the percentage coverage of land use in 5×5 Km areas inside the Negeri Sembilan state. Table 2 presents the summary of these three datasets.

3 Results and discussion

The main results are summarized and presented in this section, while a full list of the ranked selected and rejected features from all datasets is reported in Appendix.

Table 3 The eigenvalue of 100 PC from Dataset 1

PC	Eigenvalue	Variance explained (%)	Cumulative explained variance (%)
PC1	1.0885e+00	2.7205e+01	27.20
PC2	7.5512e−01	1.8872e+01	46.08
PC3	3.2765e−01	8.1888e+00	54.27
PC4	2.8520e−01	7.1276e+00	61.39
PC5	1.8382e−01	4.5941e+00	65.99
PC6	1.5967e−01	3.9905e+00	69.98
PC7	1.3266e−01	3.3154e+00	73.29
PC8	1.0958e−01	2.7387e+00	76.03
PC9	7.8591e−02	1.9641e+00	77.99
PC10	7.6475e−02	1.9113e+00	79.90
PC11	5.6561e−02	1.4136e+00	81.32
PC12	5.4966e−02	1.3737e+00	82.69
PC13	5.2167e−02	1.3038e+00	83.99
PC14	4.7305e−02	1.1822e+00	85.18
…	…	…	…
PC97	5.8702e−05	1.4671e−03	99.99
PC98	2.9117e−05	7.2769e−04	99.99
PC99	2.8408e−05	7.0997e−04	99.99
PC100	2.0795e−05	5.1970e−04	100.00

3.1 Selected features by Targeted PCA

3.1.1 Dataset 1: Communities and crime dataset

Table 3 shows that the first 14 principal components (PC) from Dataset 1 have explained variance exceeding 1%. The cumulative proportion of the first 14 PC is 85%.

Among the first 14 PCs, PC1 and PC5 were chosen as the reference PC, as both PC consisted of higher $C_{j,p}$ (3.9184% and 6.5057%) of the dependent variable ('ViolentCrimesPerPop') compared to that of other variables (Table 8). Based on these two PC, features with $C_{j,p}$ above the cutoff 1% were selected as features associated with the dependent variables. Therefore, the first 50 highest-ranked features from PC1 and the first 27 highest from PC5 were selected. The selected features may be determined by PC1 or PC5 or both. For example, the variable names 'PopDens,' 'PctVacMore6Mos,' 'PctSpeakEnglOnly,' 'PctSameState85,' and 'PctSameHouse85' were chosen because these variables contributed $C_{j,p}$ more than 1% for both PCs. Overall, 70 variables out of 100 were selected for a high association with the dependent variable ('ViolentCrimesPerPop').

Several important keys exist in predicting the total number of crimes [27]. They are divided into four major groups: socioeconomic disparities, education and literacy levels, family structure, and drug abuse or addiction. Based on our analysis, socioeconomic variables such as poverty rates (NumUnderPov), income inequality (medIncome), and unemployment rates (PctUnemployed) were found to correlate with higher crime rates. As discussed in the referenced study [27], individuals in economically disadvantaged areas often face limited opportunities and reduced access to education, healthcare, and employment, leading to

Table 4 The eigenvalue of 385 principal components from Dataset 2

PC	Eigenvalue	Variance explained (%)	Cumulative explained variance (%)
PC1	5.0091e+02	9.4009e+01	94.01
PC2	7.6621e+00	1.4380e+00	95.44
PC3	2.5762e+00	4.8350e−01	95.93
…	…	…	…
PC384	4.5027e−30	8.4505e−31	99.99
PC385	4.5027e−30	8.4505e−31	100.00

frustration, desperation, and higher rates of criminal behavior. Besides, areas with low educational attainment and high illiteracy rates often experience higher crime rates. Inadequate access to quality education can limit individuals' prospects, leading to a higher probability of involvement in violent crime. Additionally, Targeted PCA ranked 'PctNotHSGrad' highly in predicting crime. This feature measures the percentage of people 25 and over that are not high school graduates. Finally, the stability of the family structure and positive social support networks significantly impacts crime rates. Broken families (TotalPctDiv), a lack of parental involvement (PctWorkMom and PctWorkMomYoungKids), and weak social networks (PctNotSpeakEnglWell) can contribute to higher crime rates as individuals may seek validation, belonging, and support from alternative sources, including criminal activities [27]. Targeted PCA also identified urbanization and the immigrant population in the city to be linked to the number of crimes [28].

3.1.2 Dataset 2: Relative location of CT slices on axial axis dataset

Table 4 shows that two PC contributed more than 1% variance percent in Dataset 2, which are PC1 and PC2. Besides, based on Table 9, both were also selected as reference PC because the $C_{j,p}$ of the dependent variable (reference) in both PCs were highest, at 0.7163% and 2.2770%, respectively.

Overall, 254 features contributed $C_{j,p}$ of more than 0.2597% (the cutoff value) and were selected as essential features to predict the relative location of CT in the human body. Of these, 183 were higher-ranked features from PC1, and 71 were from PC2. The Targeted PCA found 149 input features from bone structure to be important in predicting the location of the CT slice. Meanwhile, only 105 features from the air inclusion group were selected. According to Furuhashict et al. (2009), the importance of histogram analysis of bone structure and air inclusion can be discussed as following [29]: (1) Bone structures play a significant role in predicting the relative location of CT slices due to their distinctive properties. Moreover, bone structures provide structural context and serve as reference points for assessing the spatial relationships between adjacent CT slices. Therefore, histograms describing bone structures are considered an important factor in predicting the relative location of CT slices on the axial axis. (2) Air inclusions, such as the lungs or air-filled cavities, also contribute to the localization of CT slices. However, air inclusions might not be as prominent as bone structures in predicting slice location, but they still provide valuable information. (3) In certain cases, particularly when dealing with thoracic or abdominal CT scans, air-filled structures can serve as reliable landmarks for determining the relative position of a slice along the axial axis. By incorporating histogram analysis of air regions, the predictive accuracy of CT slice localization can be further improved.

Table 5 The eigenvalue of 215 principal components from Dataset 3

PC	Eigenvalue	Variance explained (%)	Cumulative explained variance (%)
PC1	3.9055e+05	3.7852e+01	37.85
PC2	3.4225e+05	3.3171e+01	71.02
PC3	2.2028e+05	2.1350e+01	92.37
PC4	2.5004e+04	2.4234e+00	94.79
PC5	1.7788e+04	1.7241e+00	96.52
PC6	8.0422e+03	7.7946e−01	97.30
…	…	…	…
PC214	3.5113e−27	3.4032e−31	100.00
PC215	3.5113e−27	3.4032e−31	100.00

3.1.3 Dataset 3: Leptospirosis incidence and land use types dataset

Based on Table 5, PC1 to PC5 were selected for investigation. However, among five PCs, only PC1 is chosen as the reference PC because this PC consists of the dependent variable ('total Leptospirosis cases') with the highest $C_{j,p}$ compared to other independent variables. The dataset resulted in only one reference PC, unlike Datasets 1 and 2, which resulted in more than one reference PC. The cutoff value for this dataset is 0.4651%. Based on this, 155 independent variables were found to be important features in predicting the total cases of Leptospirosis (Table 10).

Ten types of land use were found to be important in determining the total number of leptospirosis in Negeri Sembilan, Malaysia, These are residential areas (LU_7), palm oil plantation (LU_4), rubber plantation (LU_23), sport complex (LU_5), roads (LU_115), oxidation pond (LU_60), schools (LU_52), monsoon drains (LU_66), bushes (LU_9), and hardware store (LU_2). Residential and roadways land uses demark the center of the human population and urbanization. The population of rats may be directly dependent on the presence of human homes, as they provide the source of food for rats via garbage [30]. Furthermore, the oxidation ponds treat wastewater received through the sewer system, where many colonies of rats are breeding and sheltering [31]. Like residential land use, a school area attracts a community of rats, as it provides a food source. *Leptospira* may infect school children through rats' urine and contact with street cats or dogs in school areas [32]. In 2016, a descriptive analysis demonstrated that Malaysian students registered the most significant cases in the country. 40% of the cases were reported to be students coming from school activities [33]. Palm and rubber plantation land uses are related to occupational exposure. Plantation workers are likely to be infected by Leptospira because they often work physically in contact with the surrounding environment. The predominant host animal in oil plantations has been shown to contribute 88.1% of the overall rat pathogenic *Leptospira* [34]. The unsafe work practices by plantation workers also catalyze this disease's infection rate. A cross-sectional study has shown that many workers have poor work practices that expose themselves to the plantation's surface soil and water environment, which is most likely contaminated with the urine of infected animals [35].

Table 6 Comparison in terms of the total number, similarities, and differences of selected features by both methods Targeted PCA and LASSO regression

	Dataset 1 ($n = 99$)	Dataset 2 ($n = 384$)	Dataset 3 (= 214)
Total selected by Targeted PCA	70	254	155
Total selected by LASSO	74	359	78
Similarities (Hamming distance)	57.58% ($n = 57/99$)	63.02% ($n = 242/384$)	42.99% ($n = 92/214$)

Table 7 Summary of performance of multi-linear regression fitted with input features selected by LASSO regression and Targeted PCA

	Dataset 1		Dataset 2		Dataset 3	
Selected features by	LASSO	Suggested method	LASSO	Suggested method	LASSO	Suggested method
Multiple R-squared	0.6908	0.6943	0.8644	0.8480	0.8883	0.8928
Adjusted R-squared	0.6781	0.6783	0.8635	0.8473	0.8863	0.8900
p value	2.2e−16	2.2e−16	2.2e−16	2.2e−16	2.2e−16	2.2e−16

3.2 Selected features by LASSO regression

Table 8 shows all the selected independent features in Dataset 1 with ranked coefficients value by using LASSO regression. Overall, 74 predictors out of 100 were identified to have a significant correlation with a dependent variable using this approach. The level of filtering achieved may be considered minimal, and theoretically, further adjustment to the value of alpha or the regulation value (L1) could be used to increase the reduction of features. This is because as the penalty value increases, the coefficients of many features will be set equal to zero. However, this regulation must be controlled because very high values will cause feature selection bias and misinterpretation during prediction [18].

According to Tables 9 and 10, LASSO regression found 359 and 78 features for Datasets 2 and 3, respectively. Dataset 3 shows the most restrictive selection where almost two-thirds of the independent variables were rejected.

3.3 Comparison between Targeted PCA and LASSO selected features

3.3.1 Similarities and differences between selected features

Table 6 shows the number of features chosen by Targeted PCA and LASSO regression for all datasets. The number of selected features using Targeted PCA was lower than that by LASSO regression with Datasets 1 and 2. The total number of selected features in Dataset 3 when using PCA is nearly double that when using LASSO regression.

According to the Hamming distance method, the Targeted PCA and LASSO regression chose 242 similar features out of 384 total features, the equivalent of 63.02% similarity, from Dataset 1. Meanwhile, Targeted PCA and LASSO regression select 57 of 99 similar features, the equivalent of 57.58% similarity, from Dataset 2. On the other hand, Dataset 3 shows the lowest similarity with only 92 out of 214 features selected by both methods. Since both

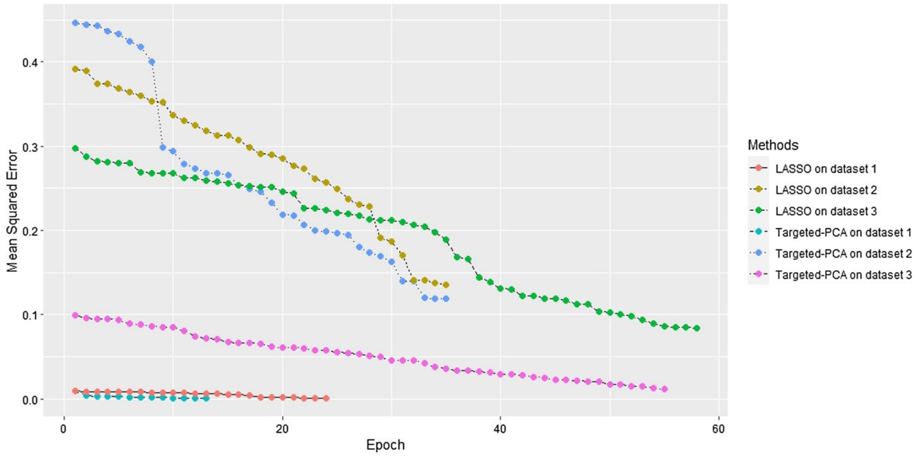

Fig. 3 Performance graph of ANN model trained by input features selected by LASSO regression and Targeted PCA

methods recorded a more significant gap in the total number of individual selected features in Dataset 3, the potential to have similar features selected by both methods was low.

All these similarities and differences may change depending on the dimensionality reduction parameters used by both methods. For example, if the study increases the cutoff of variance percent from 1% to 5%, PC 5 might be not selected as a reference PC since the variance percent is 4.6%, which is lower than the threshold value. In this case, the analysis would reject almost 28 selected features. The same goes for LASSO regression. In conclusion, the study found that both methods share more than 50% similarity of independent variables for Datasets 1 and 2. Meanwhile, Dataset 3 has less than 50% similarity of independent variables, which means there is a significant difference in the selected and rejected features by Targeted PCA and LASSO regression.

3.3.2 Prediction performance on linear and nonlinear model

This section presents the impact of selected features in identifying the linearity and nonlinearity between input and output prediction tasks.

Table 7 shows the summary of the trained and tested multiple linear regression (MLR) model, which used the selected input from all three datasets by both approaches. Both methods produce the same p value values that are lower than 0.05 for all datasets. In addition, model prediction performances when using selected features from LASSO regression and Targeted PCA are not significantly different for all datasets. The difference for multiple R^2 and adjusted R^2 was less than 0.02.

Figure 3 shows the tested ANN performance at multiple epochs comparing the different sets of selected features for all datasets. The model trained using the selected input in Dataset 1 from Targeted PCA produced a slightly higher starting error than the model trained with the input by LASSO regression. However, it recorded a drastic reduction in error for the second epoch, finally converging at epoch 13. In contrast, with the selected features by LASSO regression, the starting error was 0.00075 lower. However, the model showed a slower convergence until epochs 11 and 17, at which there are significant changes in the next epoch's error reduction. Finally, the model converged at epoch 24, with an error higher than the model trained by selected features by Targeted PCA.

With Dataset 2, the results were similar. The selected features from the Targeted PCA showed a larger error of 0.4463 at the beginning, while the model with input from LASSO had a lesser error of 0.3914. However, the condition changed when the model with the Targeted PCA performed very aggressive training when the model demonstrated a significant reduction repeatedly, especially between epochs 8 and 9. The error changed from 0.4001 to 0.2991. However, the error in the model with input from the LASSO regression gradually decreased until epoch 28, when the error started showing a significant decrease from 0.2286 to 0.1909. Figure 3 also shows both models converged at the same number of epochs, which is 35, but the model with input from the Targeted PCA produced a better performance than the model that was trained with input from LASSO with final MSE values of 0.1186 and 0.1355, respectively.

Dataset 3 shows the ANN model trained with the input from the Targeted PCA performed better than the model with selected features from LASSO from the beginning until the last epoch. The model with the Targeted PCA produced 0.099 MSE, while the model with LASSO regression produced 0.2973 MSE at the beginning epoch. Then, both models gradually decreased the error for the following epoch. However, the model with the input from Targeted PCA converged much faster at epoch 55 with a final error at 0.0113. Meanwhile, the model with input from LASSO regression converged with additional epochs at epoch 58, and the final error was higher at 0.0844. All trained models seemed to converge using the first rule of the early stopping algorithm, whereby the training stopped at specific epochs when the difference in the loss between previous and current epochs is lower than 0.001.

In conclusion, both methods have shown a good ability to capture the relationship between the input and output in the dataset when linearity was assumed through multi-linear regression. However, the ANN model trained faster and had better performance (lower error) with the selected features from the Targeted PCA. The selected features from the Targeted PCA provided more informative nonlinear connections between the input–output than those from the LASSO regression. Besides, the LASSO regression technique may have been underfitted the linear fitted model. To overcome the nonlinearity problem in the LASSO regression technique, previous researchers have used other LASSO variants applied for the nonlinear feature problem such as Least Absolute Shrinkage and Selection Operator-Neural Network (LassoNET) and Least Absolute Shrinkage and Selection Operator-Multi-Layer Perceptron (LassoMLP). However, these two methods are embedded feature selection methods that may not perform well with other classifiers [36]. Meanwhile, other traditional nonlinear feature selection methods such as distance correlation, Hilbert–Schmidt Information Criterion, and Hoeffding's test have suffered from ignoring the joint contribution of features in predicting the target data [37]. None of the above studies was aimed to assess regression performance using the selected features.

4 Conclusion

This study proposed a new approach using PCA for feature selection. It identified and ranked the important features based on the independent variable's connection to the selected principal component. The methodology was tested for three different datasets from different fields to ensure its robustness. The study found 2 out of 3 datasets to have above 50% similarities in selected features when compared to features selected using LASSO regression. On the other hand, the results of the feature selection indicate that the Targeted PCA performed efficiently in capturing both linear and nonlinearity patterns in the dataset in prediction tasks. The Targeted PCA produced a faster convergence and better performance in the ANN training.

The Targeted PCA method has a limitation in that it focuses on selecting the features in the dataset that belong to the reference PC with a particular threshold value. It only considers the PC with a high eigenvalue (variance explained percentage higher than 1%) and $C_{j,p}$ value from the dependent variable. Consequently, it may not be applicable to datasets that have their dependent variable with a low $C_{j,p}$ value in high-ranked PCs. To address this, future studies could investigate the effectiveness of different methods of feature transformation of the original dataset prior to the PCA.

Acknowledgements This work was supported by a Grant (No.: NEWTON/1/2018/WAB05/UPM/1) from the UK Natural Environment Council and the Ministry of Education Malaysia under the Understanding of the Impacts of Hydrometeorological Hazards program for their collaboration in research planning and development. We also acknowledge the Negeri Sembilan Town and Rural Planning Department and International Institute for Applied System Analysis for providing environmental data for free. We would like to thank the Director General of Health Malaysia for his permission to publish this article.

Author Contributions FR, ZZ, and AJ conceived the research. FR processed all the data, performed the analyses, and interpreted the results. FR and ZZ wrote the manuscript with contributions from all authors.

Data availability Datasets 1 and 2 can be retrieved from the open-source UCI Machine Learning Repository collection. Meanwhile, Dataset 3 is subject to the following licenses/restrictions: The datasets are owned by multiple government agencies and have sharing restrictions. Requests to access these datasets and codes should be directed to fariqrahmat94@gmail.com.

Declarations

Conflict of interest The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Ethical approval Ethical approval for this study was obtained from the Medical Research and Ethics Committee (MREC), Ministry of Health Malaysia (NMRR-19-4115-47702).

Appendix A Table of Dataset 1

See Table 8.

Table 8 The rank of all features in Dataset 1 is based on the Targeted PCA and LASSO regression

No	Features ID	Targeted PCA (PCA)				LASSO	
		PC1	Rank	PC5	Rank	Coefficient	Rank
1	whitePerCap	0.0119	65	0.4707	45	0.1395	18
2	TotalPctDiv	1.8111	24	6.4249	1	Reject	NA
3	RentMedian	1.7682	27	0.0189	91	Reject	NA
4	RentLowQ	0.0000	73	6.3737	2	0.2122	5
5	RentHighQ	0.0000	74	1.8447	16	Reject	NA
6	racePctWhite	0.1169	53	2.1837	12	0.0256	59
7	racePctHisp	0.0621	55	0.918	31	0.0501	36
8	racepctblack	0.1232	52	6.1622	3	0.1984	6
9	racePctAsian	0.1124	54	0.9314	30	Reject	NA
10	PopDens	1.0203	48	3.4117	8	0.0019	73
11	PersPerRent -OccHous	0.0000	75	0.0701	81	0.0755	27
12	PersPerOwn -OccHous	0.0000	76	1.6102	19	0.1171	19
13	PersPerOccup -Hous	3.1971	4	0.0738	80	0.3197	1
14	PersPerFam	1.2178	32	0.1817	61	Reject	NA
15	perCapInc	1.1338	39	0.4779	44	Reject	NA
16	PctYoungKid -s2Par	0.0000	77	1.4224	21	0.035	50
17	pctWWage	0.0284	61	1.8476	15	0.1629	11
18	pctWSocSec	1.1904	36	0.5374	41	0.0495	38
19	pctWRetire	2.2011	20	0.5238	43	0.0864	23
20	pctWPubAsst	2.6608	10	0.5268	42	Reject	NA
21	PctWorkMomYoungKids	1.8857	23	0.1268	65	0.0263	57
22	PctWorkMom	0.0000	78	5.8458	4	0.1468	14
23	PctWOFullPlumb	1.1991	35	0.0297	87	0.0076	68
24	pctWInvInc	0.0263	63	0.545	40	0.1508	13
25	pctWFarmSelf	0.028	62	0.5799	39	0.0367	48
26	PctVacMore6Mos	1.0159	49	3.3602	9	0.0627	32
27	PctVacantBoarded	2.5184	15	0.0402	85	0.0504	35
28	PctUsePubTrans	0.0000	79	1.4604	20	0.034	51
29	pctUrban	3.633	1	0.6478	37	0.039	46
30	PctUnemployed	1.211	33	0.243	55	0.0088	66
31	PctTeen2Par	1.0456	44	0.1346	64	Reject	NA
32	PctSpeakEnglOnly	1.0451	45	4.636	5	0.0026	72
33	PctSameState85	1.026	47	3.5279	7	Reject	NA
34	PctSameHouse85	1.0077	50	2.5826	10	Reject	NA
35	PctSameCity85	0.0000	80	1.1163	27	0.0291	55
36	PctRecImmig8	2.874	7	0.0989	75	0.0253	60
37	PctRecImmig5	2.5398	13	0.1001	74	Reject	NA
38	PctRecImmig10	0.0000	81	0.0967	76	Reject	NA
39	PctRecentImmig	0.0000	82	0.1008	73	Reject	NA
40	PctPopUnderPov	0.0082	66	2.4247	11	0.1468	15
41	PctPersOwnOccup	0.0000	83	0.0537	82	0.0797	25

Table 8 continued

No	Features ID	Targeted PCA (PCA)				LASSO	
		PC1	Rank	PC5	Rank	Coefficient	Rank
42	PctPersDenseHous	0.0000	84	0.049	83	0.1637	10
43	PctOccupMgmtProf	2.1481	21	0.1963	59	0.0329	52
44	PctOccupManu	0.0004	70	0.2052	58	0.0483	40
45	PctNotSpeakEnglWell	2.3288	18	0.0963	77	0.0793	26
46	PctNotHSGrad	2.6658	8	0.2528	53	0.0002	74
47	PctLess9thGrade	0.0058	67	0.2734	52	0.0549	33
48	PctLargHouseOccup	0.0000	85	0.0834	79	0.044	45
49	PctLargHouseFam	0.0000	86	0.0913	78	0.0639	31
50	PctKids2Par	0.0000	87	0.1347	63	0.2791	2
51	PctImmig -Recent	1.1711	38	0.115	69	0.0045	69
52	PctImmigRec8	2.349	17	0.1016	71	0.0026	71
53	PctImmigRec5	2.4864	16	0.115	70	Reject	NA
54	PctImmigRec10	0.0000	88	0.1012	72	Reject	NA
55	PctIlleg	3.4528	3	0.1192	67	0.1456	17
56	PctHousOwn -Occ	1.5261	30	0.0414	84	Reject	NA
57	PctHousOccup	0.0000	89	1.3699	22	0.0518	34
58	PctHousNo -Phone	2.2553	19	0.0309	86	0.0101	64
59	PctHousLess3 -BR	0.0000	90	1.6711	18	0.0676	30
60	PctForeignBorn	0.0000	91	0.0000	98	0.0736	28
61	PctFam2Par	2.5551	11	0.1661	62	Reject	NA
62	PctEmplProf -Serv	1.1074	41	0.2086	57	Reject	NA
63	PctEmploy	0.0031	69	1.2848	23	0.1465	16
64	PctEmplManu	2.6608	9	0.2281	56	0.0457	43
65	PctBSorMore	0.0034	68	0.2506	54	0.0371	47
66	PctBornSame -State	0.0000	92	1.1262	25	Reject	NA
67	OwnOccMed -Val	1.7738	26	0.0264	89	Reject	NA
68	OwnOccLow -Quart	0.0000	93	0.0268	88	0.0477	41
69	OwnOccHi -Quart	1.1105	40	0.0212	90	0.0045	70
70	OtherPerCap	1.7854	25	0.298	49	0.0443	44
71	NumUnderPov	1.7057	28	0.2901	51	Reject	NA
72	NumStreet	3.5503	2	0.0022	97	0.189	7
73	NumInShelters	1.1782	37	0.0039	96	0.0999	22
74	NumImmig	1.9818	22	0.1186	68	0.1107	20
75	NumIlleg	1.1024	42	0.121	66	0.0732	29
76	numbUrban	0.0337	60	0.6854	36	0.0462	42
77	MedYrHous -Built	0.0000	94	2.1833	13	0.0084	67
78	MedRentPct -HousInc	0.0000	95	0.0147	93	0.0494	39
79	MedRent	0.0000	96	0.0176	92	0.239	4
80	MedOwnCost -PctIncNoMtg	3.1428	5	0.0069	95	0.0845	24
81	MedOwnCost -PctInc	0.0000	97	0.0139	94	0.035	49
82	MedNumBR	1.0395	46	4.3088	6	0.0121	63

Table 8 continued

No	Features ID	Targeted PCA (PCA)				LASSO	
		PC1	Rank	PC5	Rank	Coefficient	Rank
83	medIncome	1.0466	43	0.5964	38	Reject	NA
84	medFamInc	0.0204	64	1.1244	26	0.0499	37
85	MalePctNev -Marr	0.0000	72	1.8142	17	0.1622	12
86	MalePctDivorce	0.0003	71	0.1835	60	0.1641	9
87	LandArea	1.234	31	0.0000	99	0.02	62
88	indianPerCap	2.5474	12	0.3339	47	0.0293	54
89	ï..population	1.6117	29	0.9718	28	Reject	NA
90	HousVacant	0.0000	98	1.2589	24	0.1641	8
91	householdsize	0.1273	51	0.9515	29	Reject	NA
92	HispPerCap	1.2097	34	0.2912	50	0.0265	56
93	FemalePctDiv	0.0000	99	1.856	14	0.1098	21
94	blackPerCap	3.0167	6	0.4548	46	0.0222	61
95	AsianPerCap	2.5198	14	0.3205	48	0.0256	58
96	agePct65up	0.0393	59	0.6989	35	0.01	65
97	agePct16t24	0.0441	58	0.75	34	Reject	NA
98	agePct12t29	0.0488	57	0.8052	33	0.2423	3
99	agePct12t21	0.0612	56	0.8648	32	0.0325	53
100	ViolentCrimesPerPop	3.9184	0	6.5057	0	Dependent	No ra

Appendix B Table of Dataset 2

See Table 9.

Table 9 The rank of all features in Dataset 2 is based on the Targeted PCA and LASSO regression

No	Features ID	Targeted PCA (PCA)				LASSO	
		PC1	Rank	PC2	Rank	Coefficient	Rank
1	reference	0.7163	0	2.277	0	Dependent	0
2	value0	0.0817	221	0.0251	201	3.0069	65
3	value1	0.0393	310	0.0201	216	1.1044	190
4	value10	0.5439	67	0	379	1.0806	192
5	value100	0.0103	368	0	369	1.5116	145
6	value101	0.0446	300	0.3275	62	1.4233	153
7	value102	0.0847	217	0.0589	154	1.3687	160
8	value103	0.0918	204	0.0307	191	0.5176	277
9	value104	0.2798	170	0.2305	74	0.6539	262
10	value105	0.2764	172	0.1751	85	1.0133	203

Table 9 continued

No	Features ID	Targeted PCA (PCA)				LASSO	
		PC1	Rank	PC2	Rank	Coefficient	Rank
11	value106	0.0681	255	0.0097	276	3.719	45
12	value107	0.0232	339	0.0822	133	0.9754	210
13	value108	0.0802	224	0.0242	202	2.7983	72
14	value109	0.5224	79	0.0003	355	0.2221	319
15	value11	0.5046	84	0.0098	273	0.2253	318
16	value110	0.5144	81	0.0511	167	2.9062	68
17	value111	0.0873	210	0.0202	215	0.8539	228
18	value112	0.2767	171	0.0585	156	0.5084	279
19	value113	0.3703	140	0.0169	233	1.175	182
20	value114	0.0111	366	0.2674	71	5.2627	27
21	value115	0.4364	115	0.0679	144	2.857	70
22	value116	0.0502	292	0.8249	49	1.9089	112
23	value117	0.0412	307	0.0774	136	0.0515	352
24	value118	0.481	98	0.1661	87	5.9019	22
25	value119	0.5728	53	0.2112	78	8.8307	10
26	value12	0.0709	248	0.0031	309	0.4073	291
27	value120	0.0169	359	0.8911	47	5.1915	28
28	value121	0.0889	206	0.0127	262	1.9616	110
29	value122	0.0661	260	0.0238	204	1.269	170
30	value123	0.3307	154	0.0165	238	1.1669	183
31	value124	0.0958	194	1.8671	14	0.2026	324
32	value125	0.282	169	0.0003	354	0.3256	304
33	value126	0.3892	134	0.0052	294	1.0743	195
34	value127	0.058	273	0.9724	44	0.4414	289
35	value128	0.4096	125	0.0059	289	2.5861	78
36	value129	0.6337	28	0.0084	281	2.4533	86
37	value13	0.0802	225	1.5724	22	0.2116	321
38	value130	0.337	153	0.2001	80	1.1259	186
39	value131	0.0384	315	0.0879	124	1.0776	193
40	value132	0.405	128	0.0165	239	4.0606	39
41	value133	0.0792	228	1.7046	19	1.7726	122
42	value134	0.4064	127	0.0116	267	0.0689	347
43	value135	0.4492	111	0.0501	168	3.505	53
44	value136	0.0875	208	0.0164	240	2.0222	107
45	value137	0.0315	324	0.0302	192	4.2849	36
46	value138	0.5011	87	0.0001	359	1.4506	151
47	value139	0.5621	57	0.1625	88	2.5039	81
48	value14	0.2848	168	0	370	0.1409	336
49	value140	0.079	230	0.0175	228	Reject	NA

Table 9 continued

No	Features ID	Targeted PCA (PCA)				LASSO	
		PC1	Rank	PC2	Rank	Coefficient	Rank
50	value141	0.3566	144	0.1002	113	0.6572	260
51	value142	0.0962	193	0.0001	366	1.6264	137
52	value143	0.2849	167	0.0668	146	1.9715	109
53	value144	0.2716	177	0.0178	226	0.9935	207
54	value145	0.0787	234	2.1297	4	1.4923	147
55	value146	0.0619	265	0.0236	206	1.496	146
56	value147	0.0237	338	0.1358	94	0.9398	215
57	value148	0.5338	73	0.128	99	1.1419	184
58	value149	0.5128	82	0.0022	320	0.4571	286
59	value15	0.0678	257	0.6163	51	0.4776	283
60	value150	0.0437	304	0.0037	303	2.4917	83
61	value151	0.0464	296	1.895	12	0.5789	268
62	value152	0.0379	317	1.927	10	1.2006	179
63	value153	0.4078	126	0.0889	123	1.6767	132
64	value154	0.0592	269	0.0647	148	0.1469	333
65	value155	0.0183	354	0.8849	48	0.2994	307
66	value156	0.0443	303	0.0141	255	1.7379	128
67	value157	0.6062	35	0.0014	332	Reject	NA
68	value158	0.4909	93	0.0073	286	1.2025	178
69	value159	0.6428	23	0.1158	105	1.6279	136
70	value16	0.0252	335	0.5308	54	0.9225	217
71	value160	0.0591	271	0.0087	280	0.4443	288
72	value161	0.047	295	0.3504	60	0.5383	272
73	value162	0.2664	181	0.0009	335	0.1326	339
74	value163	0.0496	294	0.0169	234	1.2966	165
75	value164	0.0955	195	1.9408	9	0.1772	329
76	value165	0.017	358	0	374	0.6336	267
77	value166	0.5336	74	0.0366	182	0.8905	223
78	value167	0.6131	33	0.0001	360	1.2774	169
79	value168	0.7045	8	0.0024	315	Reject	NA
80	value169	0.6061	36	0.0163	241	0.0034	358
81	value17	0.0445	302	0.0025	314	0.0837	345
82	value170	0.0544	281	0.0867	127	2.3877	91
83	value171	0.3873	136	0.0232	207	0.0903	344
84	value172	0.068	256	1.5656	24	0.0609	349
85	value173	0.25	183	0.2331	72	0.5169	278
86	value174	0.4726	104	0.0749	139	2.4177	88
87	value175	0.0502	293	0.002	323	1.2995	164
88	value176	0.6164	30	0.0124	264	3.7979	42
89	value177	0.02	351	0.017	232	0.8306	233
90	value178	0.5921	46	0.0038	300	10.0854	8

Table 9 continued

No	Features ID	Targeted PCA (PCA)				LASSO	
		PC1	Rank	PC2	Rank	Coefficient	Rank
91	value179	0.6932	11	0.0023	318	Reject	NA
92	value18	0.0032	380	0.0003	351	2.6263	77
93	value180	0.4778	101	0.0401	174	2.157	101
94	value181	0.0144	361	0.0932	121	0.2917	309
95	value182	0.0308	327	0.0877	125	2.1938	97
96	value183	0.0656	261	1.3121	33	1.2674	171
97	value184	0.0188	353	0.2325	73	0.1676	332
98	value185	0.0584	272	0.0006	340	1.4911	148
99	value186	0.0314	325	0.002	324	Reject	NA
100	value187	0.0825	218	0.0637	150	3.7113	46
101	value188	0.5869	47	0.0079	284	6.501	18
102	value189	0.7078	5	0.0349	184	Reject	NA
103	value19	0.0932	201	0.1053	112	2.9385	67
104	value190	0.0518	287	0.0022	321	0.0166	357
105	value191	0.4419	114	0.0001	367	2.5266	79
106	value192	0.0227	342	0.0566	157	Reject	NA
107	value193	0.3116	161	0.005	296	1.1033	191
108	value194	0.0544	282	0.0001	362	0.2047	323
109	value195	0.4661	107	0.0055	292	1.8613	115
110	value196	0.5708	55	0.0016	330	1.6472	135
111	value197	0.6341	27	0.0732	140	0.8491	229
112	value198	0.633	29	0.0028	310	Reject	NA
113	value199	0.5555	60	0	376	5.2954	26
114	value2	0.0985	186	0.0173	229	0.8926	222
115	value20	0.0646	262	0.1305	98	1.2823	168
116	value200	0.0001	384	0	380	0.959	211
117	value201	0.0808	222	1.2332	37	0.5308	275
118	value202	0.0414	306	1.6461	21	1.6007	139
119	value203	0.0607	267	0.0208	213	1.0186	201
120	value204	0.4157	122	0	381	0.2676	313
121	value205	0.4209	119	0.0083	282	0.3117	306
122	value206	0.4019	131	0.0126	263	1.6017	138
123	value207	0.0793	227	0.0051	295	1.0522	198
124	value208	0.5448	65	0.0262	196	0.3285	303
125	value209	0.6679	19	0.0374	179	Reject	NA
126	value21	0.0087	373	0.3455	61	0.5344	273
127	value210	0.0859	214	0.0363	183	1.5982	140
128	value211	0.0247	336	0.0008	337	3.0543	63
129	value212	0.022	345	0.2222	75	2.4911	84
130	value213	0.3484	148	0.0075	285	1.0299	200
131	value214	0.0921	203	1.1394	39	0.6651	257

Table 9 continued

No	Features ID	Targeted PCA (PCA)				LASSO	
		PC1	Rank	PC2	Rank	Coefficient	Rank
132	value215	0.2724	176	0.1431	92	1.8045	120
133	value216	0.2701	178	0.0107	270	0.1736	330
134	value217	0.5452	64	0.0158	246	1.8864	113
135	value218	0.0577	274	0.015	250	3.0474	64
136	value219	0.5428	68	0.0049	297	0.0593	350
137	value22	0.0059	377	1.5302	29	1.6763	133
138	value220	0.4704	105	0.0031	308	2.1164	102
139	value221	0.3128	159	0.1219	102	0.2723	312
140	value222	0.073	243	0.6069	52	3.3075	58
141	value223	0.0805	223	1.5632	25	0.9218	218
142	value224	0.0982	189	0.9347	46	2.0028	108
143	value225	0.037	319	1.8287	16	3.1214	62
144	value226	0.4339	117	0.014	256	2.7281	75
145	value227	0.2733	175	0.0551	161	3.5395	51
146	value228	0.4728	103	0.0254	199	3.4414	54
147	value229	0.5046	85	0.0339	186	4.2807	37
148	value23	0.0711	247	0.0191	220	1.2615	172
149	value230	0.0566	275	0.0056	291	1.9424	111
150	value231	0.0514	288	0.0001	363	0.6717	255
151	value232	0.3387	151	0.0097	275	2.3601	92
152	value233	0.0548	279	0.4474	57	0.5049	280
153	value234	0.3569	143	0.0664	147	0.5474	271
154	value235	0.0388	311	1.125	40	0.342	301
155	value236	0.4488	112	0.0003	350	1.34	162
156	value237	0.0149	360	0.0998	114	3.2329	60
157	value238	0.0854	215	0	382	2.1888	98
158	value239	0.5571	59	0.0438	173	0.8163	235
159	value24	0.3614	142	0.0163	242	0.8763	225
160	value240	0.0882	207	1.5517	26	0.2047	322
161	value241	0.0905	205	0.7615	50	0.1857	327
162	value242	0.0523	285	1.7757	17	2.4701	85
163	value243	0.2967	164	0.1531	91	0.0352	355
164	value244	0.0075	374	0.304	64	0.729	245
165	value245	0.0972	191	0.0513	166	Reject	NA
166	value246	0.698	10	0.0367	181	14.7552	4
167	value247	0.6925	12	0.0397	175	11.1959	7
168	value248	0.0387	312	0.0028	311	0.7841	239
169	value249	0.0446	301	1.9959	7	1.0569	197
170	value25	0.0868	211	2.1138	5	0.6671	256
171	value250	0.4484	113	0.0271	195	0.7023	252
172	value251	0.4532	110	0.0986	117	1.8638	114

Table 9 continued

No	Features ID	Targeted PCA (PCA)				LASSO	
		PC1	Rank	PC2	Rank	Coefficient	Rank
173	value252	0.0703	249	0.0519	165	1.5383	144
174	value253	0.0261	333	1.3004	34	0.7468	242
175	value254	0.69	13	0.0595	153	19.7067	2
176	value255	0.7046	7	0.2073	79	2.0547	104
177	value256	0.0535	284	2.2443	1	Reject	NA
178	value257	0.4633	108	0.0391	177	0.7719	240
179	value258	0.3505	146	0.0004	343	0.0652	348
180	value259	0.2668	180	0.0015	331	0.3884	296
181	value26	0.3182	157	0.0143	254	0.7543	241
182	value260	0.4132	124	0.015	251	0.1418	335
183	value261	0.0929	202	0.1741	86	0.1381	338
184	value262	0.6386	25	0.1127	108	7.0967	16
185	value263	0.6419	24	0.0183	224	1.473	149
186	value264	0.3415	150	0.0211	212	1.4501	152
187	value265	0.3021	163	0.0053	293	0.7035	251
188	value266	0.3923	132	0.023	209	1.2958	166
189	value267	0.3617	141	0.085	129	0.3233	305
190	value268	0.2748	173	0.0002	358	1.1069	189
191	value269	0.4936	92	0.0638	149	1.6782	131
192	value27	0.5784	50	0.0954	120	8.7888	11
193	value270	0.0268	332	0.0539	162	1.3995	157
194	value271	0.5955	45	0.0031	306	Reject	NA
195	value272	0.0309	326	0.0027	312	1.1789	181
196	value273	0.423	118	0	383	4.8132	31
197	value274	0.2902	166	0.0081	283	2.0279	105
198	value275	0.4197	120	0	375	0.6584	259
199	value276	0.0642	263	0.0199	217	Reject	NA
200	value277	0.0368	320	1.0966	42	3.6159	50
201	value278	0.529	76	0.0396	176	2.2549	96
202	value279	0.7148	1	0.0181	225	4.5911	34
203	value28	0.0554	278	0.0166	237	11.3251	6
204	value280	0.5252	77	0.0195	218	0.913	221
205	value281	0.3862	137	0.2209	76	1.412	155
206	value282	0.2691	179	0.004	299	0.534	274
207	value283	0.4167	121	0.0907	122	0.7264	246
208	value284	0.0407	309	0.0127	261	0.2499	315
209	value285	0.4787	100	0.0187	222	0.1791	328
210	value286	0.6699	18	0.0016	328	0.6475	265
211	value287	0.603	41	0.0333	188	7.5118	14
212	value288	0.0792	229	0.1408	93	1.7566	124
213	value289	0.3121	160	0.0167	236	0.808	237

Table 9 continued

No	Features ID	Targeted PCA (PCA)				LASSO	
		PC1	Rank	PC2	Rank	Coefficient	Rank
214	value29	0.5161	80	0.009	279	3.8808	40
215	value290	0.0194	352	1.6637	20	0.0394	354
216	value291	0.0937	197	0.2777	68	0.9213	220
217	value292	0.0183	355	0.2703	70	0.1859	326
218	value293	0.4792	99	0.0214	211	2.1732	100
219	value294	0.0006	383	0.0007	338	4.0743	38
220	value295	0.5983	43	0.0731	141	0.8106	236
221	value296	0.0051	378	0.0229	210	0.9986	205
222	value297	0.0749	240	1.5509	27	1.4109	156
223	value298	0.0027	382	0.5687	53	0.4653	284
224	value299	0.0747	241	2.0021	6	0.3527	300
225	value3	0.4032	129	0.0598	151	6.5645	17
226	value30	0.4901	94	0.0001	361	0.6973	254
227	value300	0.0821	220	1.149	38	1.2395	174
228	value301	0.0334	323	1.9186	11	0.1873	325
229	value302	0.6032	40	0	378	0.4386	290
230	value303	0.0411	308	0.0019	325	6.4688	19
231	value304	0.0514	289	0.0288	193	0.1449	334
232	value305	0.0965	192	0.2732	69	0.2313	317
233	value306	0.296	165	0.0009	336	0.9257	216
234	value307	0.0737	242	0.038	178	1.7526	125
235	value308	0.0063	376	0.4365	59	0.1384	337
236	value309	0.0244	337	2.1764	3	Reject	NA
237	value31	0.0632	264	2.1807	2	0.9489	214
238	value310	0.6382	26	0.0022	322	7.4751	15
239	value311	0.6989	9	0.0024	317	8.3131	12
240	value312	0.4867	97	0.1146	107	0.2473	316
241	value313	0.4137	123	0.0851	128	0.8871	224
242	value314	0.0201	350	1.8532	15	2.1839	99
243	value315	0.0592	270	0.2792	67	1.0758	194
244	value316	0.4884	96	0.0337	187	1.2401	173
245	value317	0.4984	89	0.0188	221	1.1328	185
246	value318	0.5442	66	0.0678	145	1.1817	180
247	value319	0.5502	63	0.0588	155	3.3358	57
248	value32	0.0935	199	1.544	28	1.062	196
249	value320	0.0457	297	0.0959	119	0.6413	266
250	value321	0.0093	371	0.1313	97	1.0511	199
251	value322	0.086	213	0.1157	106	1.2842	167

Table 9 continued

No	Features ID	Targeted PCA (PCA)				LASSO	
		PC1	Rank	PC2	Rank	Coefficient	Rank
252	value323	0.0217	348	0.0787	135	2.8141	71
253	value324	0.0098	370	0.0003	349	0.8329	232
254	value325	0.0721	246	0.0017	327	0.3912	294
255	value326	0.0295	328	0.0153	248	0.3915	293
256	value327	0.5503	62	0.0098	274	3.6542	48
257	value328	0.0765	237	1.29	35	0.354	298
258	value329	0.038	316	0.0532	163	0.9536	212
259	value33	0.0457	298	0.0326	189	0.5623	270
260	value330	0.4996	88	0.107	111	0.7413	243
261	value331	0.0092	372	1.0194	43	0.283	311
262	value332	0.3154	158	0.0825	130	0.124	342
263	value333	0.5408	70	0.1602	89	0.6492	264
264	value334	0.6025	42	0	368	1.7113	130
265	value335	0.0226	343	0.097	118	0.8714	226
266	value336	0.0385	314	0.1788	83	0.7007	253
267	value337	0.0945	196	1.4868	30	Reject	NA
268	value338	0.07	251	0.0172	231	0.9755	209
269	value339	0.5044	86	0.0144	253	1.4714	150
270	value34	0.0281	329	1.1113	41	0.8253	234
271	value340	0.0677	258	0.0177	227	2.3904	90
272	value341	0.0219	346	0.0056	290	5.1795	29
273	value342	0.708	4	0.0095	278	4.4498	35
274	value343	0.023	341	0.0005	342	0.7091	250
275	value344	0.0355	322	0.0161	244	0.8343	231
276	value345	0.338	152	0.0868	126	0.1081	343
277	value346	0.327	155	0.0254	200	0.3538	299
278	value347	0.5674	56	0.0163	243	2.5192	80
279	value348	0.5814	48	0.0006	339	0.078	346
280	value349	0.6551	20	0.1262	100	2.2792	93
281	value35	0.3504	147	0.134	96	3.3948	55
282	value350	0.6873	14	0.0024	316	Reject	NA
283	value351	0.713	2	0.0068	287	Reject	NA
284	value352	0.3846	138	0	371	0.2598	314
285	value353	0.3822	139	0.0003	346	0.2219	320
286	value354	0.0689	254	0.0013	333	0.1712	331
287	value355	0.3874	135	0.1222	101	0.4816	281
288	value356	0.052	286	0.0491	169	0.8057	238
289	value357	0.6045	39	0.0005	341	0.9533	213
290	value358	0.6147	31	0.1798	82	6.1181	20
291	value359	0.6071	34	0.024	203	2.4134	89
292	value36	0.0364	321	1.5701	23	3.6906	47

Table 9 continued

No	Features ID	Targeted PCA (PCA)				LASSO	
		PC1	Rank	PC2	Rank	Coefficient	Rank
293	value360	0.0378	318	0.3036	65	0.1326	340
294	value361	0.3516	145	0.0004	344	0.5177	276
295	value362	0.4888	95	0.0466	171	3.5244	52
296	value363	0.0937	198	0.4649	56	1.2345	176
297	value364	0.4768	102	0.0804	134	1.8322	116
298	value365	0.4962	90	0.0038	301	0.0028	359
299	value366	0.5804	49	0.0003	347	1.82	118
300	value367	0.6726	16	0.0067	288	1.5765	141
301	value368	0.5372	71	0.1109	109	0.4599	285
302	value369	0.0387	313	0.5065	55	0.289	310
303	value37	0.5719	54	0.0147	252	Reject	NA
304	value370	0.4354	116	0.0696	143	0.8485	230
305	value371	0.0138	362	0.0368	180	0.8567	227
306	value372	0.0752	239	0.2949	66	0.7151	248
307	value373	0.0995	184	1.3434	31	0.6497	263
308	value374	0.018	357	0.0756	137	5.7319	24
309	value375	0.079	231	0.0001	365	2.7896	73
310	value376	0.0231	340	0.0003	353	0.1305	341
311	value377	0.0983	188	0.3105	63	0.4014	292
312	value378	0.3233	156	0.1778	84	1.2377	175
313	value379	0.028	330	0.0596	152	0.9776	208
314	value38	0.5059	83	0.0992	115	4.6245	33
315	value380	0.0761	238	0.0004	345	0.655	261
316	value381	0.47	106	0.0003	348	1.6577	134
317	value382	0.051	291	0.1902	81	4.8572	30
318	value383	0.0183	356	0.0003	352	5.8625	23
319	value39	0.0613	266	0.0044	298	Reject	NA
320	value4	0.2648	182	0.0129	260	6.1084	21
321	value40	0.0563	276	0.0237	205	3.3036	59
322	value41	0.0776	236	0.0022	319	0.9939	206
323	value42	0.0788	233	1.8894	13	2.7802	74
324	value43	0.0724	245	0.0002	356	Reject	NA
325	value44	0.0985	187	0	373	2.4309	87
326	value45	0.4552	109	0.0521	164	1.7445	127
327	value46	0.043	305	0.1084	110	0.0236	356
328	value47	0.5367	72	0.0123	265	1.4164	154
329	value48	0.0111	367	0.0136	257	53.3376	1
330	value49	0.5783	51	0.0018	326	2.2742	94

Table 9 continued

No	Features ID	Targeted PCA (PCA)				LASSO	
		PC1	Rank	PC2	Rank	Coefficient	Rank
331	value5	0.0449	299	0.0708	142	3.3714	56
332	value50	0.0226	344	0.0559	160	2.902	69
333	value51	0.4943	91	0.0002	357	1.357	161
334	value52	0.0072	375	0.957	45	1.3293	163
335	value53	0.0137	363	0.0001	364	1.3844	158
336	value54	0.0114	365	0.0316	190	0.2974	308
337	value55	0.0124	364	0.0151	249	1.8316	117
338	value56	0.5983	44	0.0129	259	0.4791	282
339	value57	0.5421	69	0.0825	131	3.8349	41
340	value58	0.6703	17	0.0111	269	0.7401	244
341	value59	0.606	37	0.0106	271	Reject	NA
342	value6	0.0975	190	0.0035	304	0.9989	204
343	value60	0.079	232	0.0563	159	0.576	269
344	value61	0.3914	133	0.0825	132	1.2176	177
345	value62	0.0604	268	1.3372	32	0.3312	302
346	value63	0.3088	162	0.099	116	2.9792	66
347	value64	0.0216	349	0.0119	266	2.5033	82
348	value65	0.576	52	0.1202	104	2.0568	103
349	value66	0.005	379	0.0097	277	2.7162	76
350	value67	0.4021	130	0.0136	258	3.7204	44
351	value68	0.7122	3	0.2192	77	Reject	NA
352	value69	0.5505	61	0	372	Reject	NA
353	value7	0.2743	174	0.0206	214	3.1934	61
354	value70	0.0274	331	0.0173	230	0.7131	249
355	value71	0.0796	226	0.0442	172	1.765	123
356	value72	0.0986	185	0.157	90	2.26	95
357	value73	0.0538	283	0.4386	58	0.4482	287
358	value74	0.0666	259	0.0038	302	1.798	121
359	value75	0.0935	200	0.016	245	1.8077	119
360	value76	0.5241	78	0	384	3.7268	43
361	value77	0.0726	244	0.1349	95	7.6988	13
362	value78	0.649	21	0.0114	268	0.0413	353
363	value79	0.6486	22	0.0346	185	Reject	NA
364	value8	0.3437	149	0.0184	223	5.6806	25
365	value80	0.0548	280	0.026	198	1.1231	188
366	value81	0.0031	381	0.0158	247	1.7483	126
367	value82	0.0874	209	0.0105	272	1.5435	143
368	value83	0.0512	290	0.0031	307	0.0593	351
369	value84	0.0868	212	0.0755	138	0.3885	295
370	value85	0.0694	252	0.001	334	1.568	142
371	value86	0.0103	369	0.0025	313	0.383	297

Table 9 continued

No	Features ID	Targeted PCA (PCA)				LASSO	
		PC1	Rank	PC2	Rank	Coefficient	Rank
372	value87	0.5333	75	0.0231	208	4.6946	32
373	value88	0.6759	15	0.0167	235	Reject	NA
374	value89	0.705	6	0.0564	158	3.6325	49
375	value9	0.0782	235	0.0488	170	1.0178	202
376	value90	0.0825	219	0.0261	197	1.3694	159
377	value91	0.0219	347	1.7144	18	2.0229	106
378	value92	0.0694	253	0	377	1.1258	187
379	value93	0.0852	216	1.2463	36	0.7158	247
380	value94	0.0261	334	0.0195	219	0.9217	219
381	value95	0.0702	250	0.0016	329	0.6626	258
382	value96	0.0557	277	1.9865	8	1.7331	129
383	value97	0.5599	58	0.0287	194	9.133	9
384	value98	0.6147	32	0.1205	103	18.5862	3
385	value99	0.6053	38	0.0035	305	13.9036	5

Appendix C Table of Dataset 3

See Table 10.

Table 10 The rank of all features in Dataset 3 is based on the Targeted PCA and LASSO regression

No	Feature ID	Targeted PCA (PCA)		LASSO	
		PC1	Rank	Coefficient	Rank
1	LU_0	0.4936	116	0.00213229	55
2	LU_1	0.5904	12	0.664464958	73
3	LU_2	1.0583	10	Reject	NA
4	LU_3	0.5868	18	0.060544005	NA
5	LU_4	1.0991	2	1.048678917	76
6	LU_5	1.0972	4	Reject	77
7	LU_6	0.0002	212	4.663301763	69
8	LU_7	1.0994	1	Reject	NA
9	LU_8	0.4724	135	0.835606956	NA
10	LU_9	1.0588	9	Reject	70
11	LU_10	0.3414	173	0.642604261	29
12	LU_11	0.3599	166	Reject	14
13	LU_12	0.5882	15	Reject	72
14	LU_13	0.4803	129	Reject	31
15	LU_14	0.3034	185	Reject	NA
16	LU_15	0.5513	65	0.1741555	NA
17	LU_16	0.0003	194	0.406984499	NA

Table 10 continued

No	Feature ID	Targeted PCA (PCA)		LASSO	
		PC1	Rank	Coefficient	Rank
18	LU_17	0.0003	199	Reject	NA
19	LU_18	0.5083	106	Reject	NA
20	LU_19	0.3992	156	Reject	NA
21	LU_20	0.5606	58	Reject	NA
22	LU_21	0.4698	140	Reject	NA
23	LU_22	0.5233	93	Reject	NA
24	LU_23	1.0979	3	0.543832757	NA
25	LU_24	0.4895	119	0.856977295	36
26	LU_25	0.3751	161	Reject	NA
27	LU_26	0.5235	92	Reject	NA
28	LU_27	0.5931	11	Reject	NA
29	LU_28	0.5161	97	0.017154528	NA
30	LU_29	0.5612	56	Reject	67
31	LU_30	0.5249	89	4.006090662	61
32	LU_31	0.3416	172	Reject	12
33	LU_32	0.5865	19	Reject	NA
34	LU_34	0.3456	171	Reject	NA
35	LU_35	0.3748	162	0.01604638	NA
36	LU_36	0.5679	47	Reject	68
37	LU_37	0.474	132	0.264527014	NA
38	LU_38	0.3412	174	Reject	45
39	LU_39	0.4709	138	0.000485054	NA
40	LU_40	0.4818	126	Reject	22
41	LU_41	0.2988	186	0.303879563	NA
42	LU_42	0.4821	125	0.001380156	41
43	LU_43	0.5899	13	0.245579513	75
44	LU_44	0.5717	38	Reject	47
45	LU_45	0.334	175	Reject	NA
46	LU_46	0.3539	169	Reject	NA
47	LU_47	0.468	146	Reject	NA
48	LU_48	0.0003	195	0.734548258	NA
49	LU_49	0.4691	143	1.02E-05	27
50	LU_50	0.4775	131	0.058103446	NA
51	LU_51	0.4926	117	0.250353693	62
52	LU_52	1.0953	7	Reject	46
53	LU_54	0.0003	203	Reject	NA
54	LU_57	0.0003	209	Reject	NA
55	LU_59	0.0004	191	0.014897198	NA
56	LU_60	1.0958	6	Reject	8

Table 10 continued

No	Feature ID	Targeted PCA (PCA)		LASSO	
		PC1	Rank	Coefficient	Rank
57	LU_61	0.0003	198	0.544445776	NA
58	LU_64	0.0003	207	0.103667061	33
59	LU_65	0.5663	48	Reject	57
60	LU_66	1.0938	8	48.05861538	NA
61	LU_67	0.4721	136	Reject	1
62	LU_68	0.503	109	Reject	NA
63	LU_69	0.4843	122	Reject	NA
64	LU_70	0.4993	111	Reject	NA
65	LU_71	0.5796	30	0.667267792	NA
66	LU_72	0.4674	148	Reject	28
67	LU_74	0.4675	147	Reject	NA
68	LU_75	0.5765	34	0.22867467	NA
69	LU_76	0.5475	74	1.201208248	49
70	LU_77	0.5233	94	Reject	19
71	LU_78	0.4729	133	Reject	NA
72	LU_79	0.5507	66	Reject	NA
73	LU_80	0.377	160	Reject	25
74	LU_81	0.4808	127	0.294027436	NA
75	LU_82	0.5493	71	Reject	42
76	LU_83	0.0003	211	23.86268745	NA
77	LU_84	0.0003	204	17.24817655	2
78	LU_85	0.0003	201	5.482019244	3
79	LU_86	0.3606	164	2.157124122	7
80	LU_87	0.5098	104	Reject	16
81	LU_88	0.4665	151	Reject	NA
82	LU_89	0.5682	45	0.012546071	NA
83	LU_90	0.0003	205	0.097611892	NA
84	LU_91	0.5781	31	0.027907954	58
85	LU_92	0.5879	16	4.378085251	64
86	LU_93	0.3276	180	Reject	9
87	LU_94	0.4728	134	Reject	NA
88	LU_95	0.4971	114	0.08131086	NA
89	LU_96	0.5154	98	Reject	60
90	LU_97	0.565	51	Reject	NA
91	LU_98	0.5774	33	Reject	NA
92	LU_99	0.3103	183	Reject	NA
93	LU_100	0.3277	179	Reject	30
94	LU_101	0.4863	121	Reject	NA
95	LU_102	0.5496	69	Reject	NA

Table 10 continued

No	Feature ID	Targeted PCA (PCA)		LASSO	
		PC1	Rank	Coefficient	Rank
96	LU_103	0.0003	200	Reject	NA
97	LU_104	0.5537	63	Reject	NA
98	LU_105	0.4806	128	Reject	NA
99	LU_106	0.4696	141	Reject	NA
100	LU_107	0.4669	150	Reject	NA
101	LU_108	0.5706	39	Reject	NA
102	LU_109	0.5248	90	2.241067936	NA
103	LU_110	0.5479	73	0.534393123	NA
104	LU_111	0.4997	110	3.234181505	35
105	LU_112	0.0005	190	1.269873289	13
106	LU_113	0.5558	61	Reject	18
107	LU_114	0.4662	152	0.025522253	NA
108	LU_115	1.096	5	Reject	65
109	LU_116	0.5633	55	Reject	NA
110	LU_117	0.5755	35	0.002058971	NA
111	LU_118	0.3842	159	0.598942764	74
112	LU_119	0.5105	102	0.00308489	32
113	LU_120	0.3278	178	Reject	NA
114	LU_121	0.584	24	Reject	NA
115	LU_122	0.0002	213	Reject	NA
116	LU_123	0.5121	100	Reject	NA
117	LU_124	0.4838	123	Reject	NA
118	LU_125	0.5274	86	0.017826069	NA
119	LU_127	0.548	72	Reject	66
120	LU_128	0.5816	26	0.286200211	NA
121	LU_129	0.4835	124	0.606223165	43
122	LU_130	0.5719	36	Reject	NA
123	LU_131	0.5682	44	Reject	NA
124	LU_132	0.5451	75	0.360278266	NA
125	LU_133	0.0003	206	0.108872398	39
126	LU_134	0.5188	96	0.269309922	56
127	LU_135	0.5404	77	Reject	44
128	LU_136	0.4705	139	Reject	NA
129	LU_137	0.0003	202	Reject	NA
130	LU_139	0.537	79	Reject	NA
131	LU_140	0.5253	87	0.004931333	NA
132	LU_143	0.4911	118	Reject	71

Table 10 continued

No	Feature ID	Targeted PCA (PCA)		LASSO	
		PC1	Rank	Coefficient	Rank
133	LU_144	0.5549	62	0.852871327	NA
134	LU_145	0.5499	67	2.181417399	24
135	LU_147	0.1437	188	Reject	15
136	LU_148	0.557	59	Reject	NA
137	LU_149	0.0003	208	Reject	NA
138	LU_150	0.4866	120	4.337239735	51
139	LU_151	0.4683	145	Reject	10
140	LU_152	0.5237	91	Reject	NA
141	LU_153	0.54	78	Reject	NA
142	LU_155	0.5499	68	Reject	NA
143	LU_156	0.569	43	Reject	NA
144	LU_157	0.5647	52	Reject	NA
145	LU_158	0.5352	81	Reject	NA
146	LU_160	0.3598	167	Reject	38
147	LU_161	0.4656	153	0.8219101	NA
148	LU_162	0.5522	64	Reject	26
149	LU_163	0.0004	193	Reject	NA
150	LU_164	0.5718	37	Reject	NA
151	LU_165	0.5804	28	Reject	NA
152	LU_166	0.565	50	8.172232949	NA
153	LU_167	0.3145	182	Reject	6
154	LU_168	0.3252	181	Reject	NA
155	LU_169	0.5069	107	Reject	NA
156	LU_170	0.0003	210	Reject	NA
157	LU_171	0.4685	144	Reject	NA
158	LU_172	0.5496	70	0.140950788	NA
159	LU_173	0.0003	197	Reject	54
160	LU_174	0.5438	76	Reject	NA
161	LU_175	0.5828	25	Reject	NA
162	LU_176	0.5659	49	Reject	NA
163	LU_177	0.525	88	1.71171669	NA
164	LU_178	0.0004	192	Reject	17
165	LU_179	0.3307	177	Reject	NA
166	LU_180	0.0003	196	0.046582821	NA
167	LU_181	0.5636	54	Reject	63
168	LU_182	0.2988	187	0.152235805	NA
169	LU_183	0.4652	155	0.177073547	52
170	LU_184	0.3066	184	Reject	50
171	LU_185	0.5197	95	Reject	NA

Table 10 continued

No	Feature ID	Targeted PCA (PCA)		LASSO	
		PC1	Rank	Coefficient	Rank
172	LU_186	0.5611	57	1.114171305	NA
173	LU_187	0.3577	168	1.25E-11	21
174	LU_189	0.3507	170	Reject	78
175	LU_190	0.0005	189	Reject	NA
176	LU_191	0.57	41	0.236043882	NA
177	LU_192	0.4671	149	Reject	48
178	LU_193	0.3988	157	Reject	NA
179	LU_194	0.5645	53	Reject	NA
180	LU_195	0	214	0.082958005	NA
181	LU_196	0.5568	60	Reject	59
182	LU_197	0.5691	42	0.145589036	NA
183	LU_198	0.5877	17	Reject	53
184	LU_199	0.5816	27	Reject	NA
185	LU_200	0.5147	99	Reject	NA
186	LU_201	0.5801	29	Reject	NA
187	LU_203	0.5848	23	0.313433025	NA
188	LU_204	0.5679	46	Reject	40
189	LU_205	0.36	165	Reject	NA
190	LU_207	0.5704	40	Reject	NA
191	LU_208	0.3701	163	17.10389786	NA
192	LU_209	0.4975	113	Reject	4
193	LU_210	0.4693	142	16.86886805	NA
194	LU_211	0.5343	82	Reject	5
195	LU_212	0.4953	115	Reject	NA
196	LU_213	0.3317	176	Reject	NA
197	LU_214	0.5774	32	Reject	NA
198	LU_216	0.4793	130	Reject	NA
199	LU_218	0.5859	21	1.165642762	NA
200	LU_219	0.5362	80	Reject	20
201	LU_220	0.3877	158	Reject	NA
202	LU_221	0.5883	14	Reject	NA
203	LU_222	0.5338	83	Reject	NA
204	LU_223	0.5101	103	Reject	NA
205	LU_224	0.5285	85	Reject	NA
206	LU_225	0.5863	20	Reject	NA
207	LU_230	0.471	137	Reject	34

Table 10 continued

No	Feature ID	Targeted PCA (PCA)		LASSO	
		PC1	Rank	Coefficient	Rank
208	LU_231	0.4984	112	0.500411721	NA
209	LU_232	0.585	22	Reject	37
210	LU_234	0.5337	84	Reject	NA
211	LU_236	0.512	101	4.126724408	NA
212	LU_237	0.5064	108	Reject	11
213	LU_238	0.5097	105	0.512689373	NA
214	LU_240	0.4655	154	Reject	23
215	casebyyear$total	1.1994	0	0.114093712	0

References

1. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3(Mar):1157–1182
2. Jain D, Singh V (2018) Feature selection and classification systems for chronic disease prediction: a review. Egypt Inf J 19(3):179–189
3. Tang J, Alelyani S, Liu H (2014) Feature selection for classification: A review. In: Algorithms and applications, data classification, p 37
4. Battiti R (1994) Using mutual information for selecting features in supervised neural net learning. IEEE Trans Neural Netw 5(4):537–550
5. Verma L, Srivastava S, Negi PC (2016) A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data. J Med Syst 40(7):1–7
6. Yu L, Liu H (2003) Feature selection for high-dimensional data: A fast correlation-based filter solution. In: Proceedings of the 20th international conference on machine learning (ICML-03), pp 856–863
7. Wosiak A, Zakrzewska D (2018) Integrating correlation-based feature selection and clustering for improved cardiovascular disease diagnosis. Complexity, 2018
8. Hall MA (1999) Correlation-based feature selection for machine learning
9. Kumar V, Minz S (2014) Feature selection: a literature review. SmartCR 4(3):211–229
10. Shahana AH, Preeja V (2016) Survey on feature subset selection for high dimensional data. In: 2016 International conference on circuit, power and computing technologies (ICCPCT), pp 1–4. IEEE
11. Song F, Guo Z, Mei D (2010) Feature selection using principal component analysis. In: 2010 international conference on system science, engineering design and manufacturing informatization, vol 1, pp 27–30. IEEE
12. Mubarak S, Darwis H, Umar F, Ilmawan LB, Anraeni S, Mude MA (2018) Feature selection of oral cyst and tumor images using principal component analysis. In: 2018 2nd east indonesia conference on computer and information technology (EIConCIT), pp 322–325. IEEE
13. Wang XD, Chen RC, Zeng ZQ, Hong CQ, Yan F (2018) Robust dimension reduction for clustering with local adaptive learning. IEEE Trans Neural Netw Learn Syst 30(3):657–669
14. Hair JF (2009) Multivariate data analysis
15. Kassambara A (2017) Practical guide to principal component methods. In: R: PCA, M (CA), FAMD, MFA, HCPC, factoextra (Vol. 2). Sthda
16. Abdi H, Williams LJ (2010) Principal component analysis. WIREs Comp Stat 2:433–459
17. Xu Y, Zhang D, Yang JY (2010) A feature extraction method for use with bimodal biometrics. Patt Recogn 43(3):1106–1115
18. Giersdorf J, Conzelmann M (2017) Analysis of feature-selection for LASSO regression models
19. Hamming R (2012) Numerical methods for scientists and engineers. Courier Corporation
20. Zhai D, Liu X, Chang H, Zhen Y, Chen X, Guo M, Gao W (2018) Parametric local multiview hamming distance metric learning. Patt Recogn 75:250–262
21. Tang M, Yu Y, Aref WG, Malluhi QM, Ouzzani M (2015) Efficient processing of hamming-distance-based similarity-search queries over MapReduce. In EDBT, pp 361–372

22. Uyanık GK, Güler N (2013) A study on multiple linear regression analysis. Proc Soc Behav Sci 106:234–240
23. Fischer MM (2015) Neural networks: a class of flexible non-linear models for regression and classification. In: Handbook of research methods and applications in economic geography. Edward Elgar Publishing
24. Rabunal JR, Dorado J (Eds.) (2006) Artificial neural networks in real-life applications. IGI Global
25. Redmond M, Baveja A (2002) A data-driven software tool for enabling cooperative information sharing among police departments. Eur J Oper Res 141(3):660–678
26. Graf F, Kriegel HP, Schubert M, Pölsterl S, Cavallaro A (2011) 2D image registration in CT images using radial image descriptors. In: International conference on medical image computing and computer-assisted intervention. Springer, Berlin, Heidelberg, pp 607–614
27. Mandalapu V, Elluri L, Vyas P, Roy N (2023) Crime prediction using machine learning and deep learning: a systematic review and future directions. IEEE Access
28. Adelman R, Reid LW, Markle G, Weiss S, Jaret C (2017) Urban crime rates and the changing face of immigration: evidence across four decades. J Ethn Crim Just 15(1):52–77
29. Furuhashi S, Abe K, Takahashi M, Aizawa T, Shizukuishi T, Sakaguchi M, Sasaki Y (2009) A computer-assisted system for diagnostic workstations: automated bone labeling for CT images. J Digit Imag 22:689–695
30. Ng M (2016) Environmental factors associated with increased rat populations: a focused practice question
31. Byers KA, Lee MJ, Patrick DM, Himsworth CG (2019) Rats about town: a systematic review of rat movement in urban ecosystems. Front Ecol Evol 7:13
32. Navarrete EJ, Rivas SB, Soriano RML (2015) Leptospirosis prevalence and associated factors in school children from Valle de Chalco-Solidaridad, State of Mexico. Int J Pediatr Res 1:8
33. Tan WL, Soelar SA, Mohd Suan MA, Hussin N, Cheah WK, Verasahib K, Goh PP (2016) Leptospirosis incidence and mortality in Malaysia. Southeast Asian J Trop Med Public Health 47(3):434–40
34. Mohamed-Hassan SN, Bahaman AR, Mutalib AR, Khairani-Bejo S (2012) Prevalence of pathogenic leptospires in rats from selected locations in peninsular Malaysia. Res J Animal Sci 6(1):12–25
35. Ridzuan J, Aziah BD, Zahiruddin WM (2016) The occupational hazard study for leptospirosis among agriculture workers. Int J Collab Res Intern Med Public Health 8:MA13–MA22
36. Lemhadri I, Ruan F, Tibshirani R (2021) Lassonet: neural networks with feature sparsity. In: International conference on artificial intelligence and statistics, pp 10–18. PMLR
37. Krakovska O, Christie G, Sixsmith A, Ester M, Moreno S (2019) Performance comparison of linear and non-linear feature selection methods for the analysis of large survey datasets. Plos one 14(3):e0213584

Zed Zulkafli is an associate professor at the Department of Civil Engineering, Universiti Putra Malaysia (UPM). Her research focusses on understanding tropical hydrometeorology and the development of process-based and data-based modelling tools to test hydrological systems' behaviours.

Fariq Rahmat is a PhD candidate at Universiti Putra Malaysia, specializing in the intersection of environment, hydrology and epidemiology. He holds a Master of Science in Control Engineering from the same university, where he is passionate about leveraging artificial intelligence to bridge the gap between these fields, intending to advance the understanding of how environmental factors affect public health. Fariq has a strong research background in statistical modelling, focusing on artificial neural networks. He is actively engaged in cutting-edge research aimed at developing innovative AI models that analyse complex datasets from hydrological and epidemiological studies. These models are designed to extract meaningful insights, predict disease outbreaks and inform evidence-based policy decisions. Fariq is a multifaceted professional, currently serving as an AI developer at Imagine AI SDN BHD, where he is actively engaged in developing generalized AI models. In this role, he focuses on leveraging their deep understanding of AI and machine learning to tackle complex challenges and enhance AI capabilities across various domains, particularly in the realm of computer vision applications.

Asnor Juraiza Ishak is an Associate Professor at the Department of Electrical and Electronic Engineering, Universiti Putra Malaysia (UPM). She is a senior member of Institute of Electrical & Electronic Engineering (IEEE). She is a Professional Technologist of Malaysia Board of Technologists. She received a Bachelor degree (BEng.) in Electrical-Mechatronic Engineering from Universiti Teknologi Malaysia in 2000. Asnor Juraiza received a M.Sc in Control & Automation Engineering from UPM, then she obtained a PhD in Electrical, Electronic & System from Universiti Kebangsaan Malaysia. Her area of interest includes; control system, image & signal processing, artificial intelligence, digital twin and rehabilitation & assistive robotics. The latest research she is focus on digital twin data driven modelling for gas turbine monitoring. Her research group focus on electromyography based artificial neural network controller for knee rehabilitation machine.

Dr. Ribhan Zafira BT Abdul Rahman is affiliated to Department of Electrical and Electronic Engineering, University Putra Malaysia. She is currently providing services as Senior Lecturer. She has authored and coauthored multiple peer-reviewed scientific papers and presented works at many national and International conferences. Her contributions have acclaimed recognition from honourable subject experts around the world. She is actively associated with different societies and academies. Her academic career is decorated with several reputed awards and funding. Her research interests include Control Modelling, Fault Detection and Diagnosis, Artificial Intelligent and Pattern Classification.

Simon De Stercke is a postdoctoral research associate in the Environmental and Water Resources Engineering section of the Department of Civil and Environmental Engineering of Imperial College London. His research has been focusing on the intersection between hydrology and epidemiology, and he also works on research impact evaluation. Simon received his PhD for studying the water-energy nexuses of Mumbai and London from an end-use perspective, using system dynamics. Before coming to Imperial College London, Simon worked at IIASA (the International Institute for Applied Systems Analysis). He holds degrees in electromechanical engineering and environmental management.

Wouter Buytaert is a professor in Hydrology and Water Resources at Imperial College London. His research focuses on the impact of environmental change on the terrestrial cycle. He applies both field methods and computational modelling to generate evidence to support water management and policy, with a specific interest in participatory approaches and knowledge co-production.

Prof Dr Wardah Tahir is a director and lecturer at Universiti Teknologi Mara, Malaysia. She received her BSc in Agricultural Engineering at Cornell University, USA, and received her MSc in Water Resources Tech. & Management from Birmingham University and was awarded with Doctor of Philosophy (Civil Engineering) majoring in Hydrometeorological Flood Forecasting at University Teknologi Mara. She is a professor of water resources and environmental systems with more than 25 years of experiences in education and academic management while pursuing as a skilled researcher in flood modeling and forecasting focusing on meteorological satellite and radar application. She has a vast experience in management through the various posts held while actively involved in research and publication with more than 80 publications including journals and books. She has also conducted numerous research projects in flood disaster management, design flood estimation, hydrodynamic flood modeling, simulation and forecasting for more than 15 years via national and institutional research grants provided. She has also involved in numerous consultancy projects with the industries in water management, water quality and sustainable environment.

Jamalludin Ab Rahman is a medical doctor with a specialization in Epidemiology and Biostatistics. He possesses a keen interest in complex survey methodology, a testament to which is his involvement in multiple national surveys throughout Malaysia. Presently, Dr. Jamalludin serves as the Dean of the Kulliyyah of Medicine at the International Islamic University Malaysia. He is the Chair for the Malaysian Medical Deans Council and is the President of the Malaysia Public Health Physicians Association.

Salwa Ibrahim is a medical officer who is currently serves as the Deputy Director of Communicable Disease Control Unit at the Negeri Sembilan State Health Department. Her job scope includes monitoring trend of infectious disease and planning infectious disease control activities.

Muhamad Ismail is a Public Health Physician who is currently served as the Deputy Director of Communicable Disease Control Unit at the Negeri Sembilan State Health Department. Her job scope includes monitoring trend of infectious disease and planning infectious disease control activities.

Authors and Affiliations

Fariq Rahmat[1] · Zed Zulkafli[2] · Asnor Juraiza Ishak[1] · Ribhan Zafira Abdul Rahman[1] · Simon De Stercke[3] · Wouter Buytaert[3] · Wardah Tahir[4] · Jamalludin Ab Rahman[5] · Salwa Ibrahim[6] · Muhamad Ismail[6]

Fariq Rahmat
fariqrahmat94@gmail.com

Asnor Juraiza Ishak
asnorji@upm.edu.my

Ribhan Zafira Abdul Rahman
ribhan@upm.edu.my

Simon De Stercke
simon.destercke@imperial.ac.uk

Wouter Buytaert
w.buytaert@imperial.ac.uk

Wardah Tahir
warda053@uitm.edu.my

Jamalludin Ab Rahman
arjamal@iium.edu.my

Salwa Ibrahim
sitislw@moh.gov.my

Muhamad Ismail
muhamad.ismail@moh.gov.my

1 Department of Electrical and Electronic Engineering, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia

2 Department of Civil Engineering, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia

3 Department of Civil and Environmental Engineering, Imperial College London, South Kensington, London SW7 2BX, UK

4 School of Civil Engineering, College of Engineering Universiti Teknologi Mara, 40450 Shah Alam, Selangor, Malaysia

5 Department of Community Medicine, Kulliyyah of Medicine, International Islamic University Malaysia, 25200 Kuantan, Pahang, Malaysia

6 Negeri Sembilan State Health Department, Ministry of Health Malaysia, 70300 Seremban, Negeri Sembilan, Malaysia