The discovery of normality of body weight using principal component analysis: a comparative study on machine learning techniques using different data pre-processing methods

M. Sornam* and M. Meharunnisa

Department of Computer Science,
University of Madras,
Chennai, 600 025, TamilNadu, India
Email: madasamy.sornam@gmail.com
Email: m.meharunnisa1987@gmail.com
*Corresponding author

Abstract: In data mining, feature selection plays an important role in finding the most important predictor variables (or features) that explain a major part of the variance of the response variable is a key to identify and build high performing models. In this proposed work, primary data is used to identify the normality/ abnormality of body weight. The missing data has been imputed by predictive mean matching (PMM) method. Efforts are made to reduce the dimensions of the data before classification using principal component analysis (PCA). The principal components obtained are passed as input to the supervised learning algorithm such as naïve Bayes, support vector machine, decision tree, K-nearest neighbour and back propagation neural network with various pre-processing methods. The effectiveness of supervised learning algorithms is evaluated, where back propagation neural network algorithm with the centre pre-processing method has shown an advantage over other classifiers.

Keywords: missing data imputation; predictive mean matching method; pre-processing techniques; principal component analysis; PCA.

Reference to this paper should be made as follows: Sornam, M. and Meharunnisa, M. (2019) 'The discovery of normality of body weight using principal component analysis: a comparative study on machine learning techniques using different data pre-processing methods', *Int. J. Knowledge Engineering and Data Mining*, Vol. 6, No. 1, pp.74–88.

Biographical notes: M. Sornam is working as an Associate Professor in Computer Science Department at University of Madras. Her areas of research include neural networks, digital image processing, machine learning, deep learning and data mining. She received her PhD from the Department of Computer Science, University of Madras.

M. Meharunnisa is pursuing her PhD from the Department of Computer Science, University of Madras. Her current research areas include machine learning and data mining.

Copyright © 2019 Inderscience Enterprises Ltd.

1 Introduction

Many developing Asian countries have a unique problem of 'double burden' whereby, at one finish of the spectrum we have obesity in kids and adolescents whereas, at the opposite finish we have malnutrition and underweight. Body weight can be classified based on body mass index (BMI). According to WHO, BMI is a simple index of weight-for-height that is generally used to classify underweight, overweight and obesity in adults. It is characterised as the weight in kilograms separated by the square of the height in meters (kg/m^2). BMI esteems are age-free and the same for both genders. Be that as it may, BMI may not relate to a similar level of heftiness in various populaces due, to some extent, to various body extents. The health risk increases with increasing BMI and the BMI grading may differ for a different population. The centres for disease control (CDC) now refer to excess weight and obesity as 'an epidemic' (Suranovic et al., 2003).

There are many factors which are responsible for determining the body weight. They are the type of food, physical activity, number of hours sitting before the television, working hours with the computer, frequency of eating the food out, type of snacks, the reason to make the people overeat, family history of obese, psychological problems, emotional problem and so on. The global classification of adult underweight, overweight, and obesity according to BMI is given in Table 1.

Table 1 The international classification of body weight

Classification	*BMI (Kg/m^2)*	*Normal/abnormal range*
Underweight	<18.50	Abnormal range
Normal	18.50–24.99	Normal range
Overweight	>=25.00	Abnormal range
Obese	>=30.00	Abnormal range

A typical approach in data science is what we call featurisation of the universe. It is that we extract and engineer all the features possible for a given problem. To give an example: in a time series problem, one could use cumulative sums, moving averages with variable window sizes, discrete state changes, average differences, etc., as features, which quickly becomes very large. The raw data may not be setup to be in the best shape of modelling. Sometimes, we need to pre-process our data in order to better present the inherent structure of the problem in our data to the modelling algorithms. Pre-processing involves techniques to transform raw data into a more understandable format (Nayak et al., 2016). Some of the commonly used pre-processing techniques are centre, scale, range and the combinations of the methods can also be used.

In this case, exploratory data analysis (EDA) is challenging and we needed to resort to alternative methods of visualising and exploring the feature space. A multivariate analysis issue could begin with a generous number of corresponded factors. Principal component analysis (PCA) is a technique for reducing a large set of variables to a small set that still contains most of the information in the large set. The aim of this paper is to reduce the dimension of data before classification and evaluate the accuracy of different classification algorithms.

One methodology is to diminish the dimensionality of the feature space and jab around in the decreased feature space. A basic technique which is ideal for this paper is PCA. The primary objective of a PCA analysis is to distinguish patterns in data; PCA means to recognise the correlation between factors. PCA reduces the dimensionality of the dataset consisting of many variables correlated with each other, either strongly or weakly, while retaining the variation present in the dataset, to the minimum extent.

This paper is organised as follows: in Section 2, the descriptions of some of the methods in practice are highlighted. In Section 3, the methods adopted are briefly explained and the details about the collection of data are explained. In Section 4, results of the proposed model are highlighted. In Section 5, the results are well discussed. The paper is concluded with its conclusion in the Section 6.

2 Literature review

With a specific end goal to clean the information from inconsistencies, distinctive procedures of data pre-processing play a major role. Abolmakarem et al. (2016) pre-processed and cleansed the data by replacing the missing value with the field mean. Also, the author used local outlier factor to represent the outlier of a record. Manek et al. (2016), performed the data pre-processing using replacing the missing values with the field average and replacing the null values with '0'. Uma and Hanumanthappa (2017), while discussing the different data pre-processing techniques in healthcare, reviewed the different methods where Expectation-maximisation algorithm and multiple imputation methods can be employed for imputation. Wu et al. (2018) proposed a model to predict diabetes mellitus, in which the missing data in the dataset are imputed by their means from the training set. Amatul et al. (2013) compared the classification accuracy of a model with the pre-processed and non-processed data. The method used to pre-process is data discretisation, where the continuous data attribute values are converted into a finite set of interval and associated with a specific data. In Patil et al. (2010), the author deleted some inappropriate and inconsistent data during data pre-processing and to develop a hybrid model, the size of the data has been reduced by deleting the irrelevant features manually from the dataset during the feature selection phase. Ahmad et al. (2011) studied the prediction accuracy of multilayer perceptron in neural network against the decision tree based algorithm such as ID3 and J48 algorithm. During data pre-processing, the dataset is transformed through generalisation process in which low level concept are transformed into higher level concept. The dataset is pre-processed using min-max normalisation. The results showed that J48 performed with higher accuracy of 89.3%. The prediction accuracy is improved to 89.7% after pruning number of times pregnant attribute. Chen et al. (2017) proposed a model, in which all the impossible and missing values are replaced by mean and then used K-means clustering algorithm to remove incorrectly classified sample. Finally, 532 out of 768 samples are left, which are then classified using J48 algorithm. The classification accuracy obtained was 90.04%. Wu et al. (2018) developed a double level algorithm using K-means to remove incorrectly clustered data. After the removal procedure, only 589 samples were left which are then classified using logistic regression classification algorithm with the accuracy of 95.42%. The missing and incorrect values are replaced by means from the training data.

Feature selection includes looking through different element subsets and assessing every one of these subsets utilising some standard (Liu and Motoda, 2012; Peña et al., 2001; Yu and Liu, 2003). Feature selection is a standout amongst the most imperative procedures of the data transformation step. It is characterised as the way toward choosing a subset of features from the element space, which is more pertinent to and instructive for the development of a model. The benefits of feature selection are numerous and identify with various parts of data analysis, for example, better perception and comprehension of information, the decrease of computational time and term of examination, and better expectation precision (Guyon and Elisseeff, 2003; Witten et al., 2016).

Kavakiotis at al. (2017) utilised the wrapper methods machine learning algorithm to assess different subsets of features and utilised them to build a predictive model since the emerging algorithm wraps the entire feature selection process. In Agarap (2018), the dataset was standardised using the mean and standard deviation of the feature. The features are then used for training the model using six machine learning algorithm such as GRU-SVM, linear regression, softmax regression, nearest neighbour search, multilayer perceptron and support vector machine. Liu et al. (2017) proposed a pre-processing method using the AdaBoost algorithm, for relabeling the mislabelled instances and to find and correct the attribute with the deviation of the instance. Zhang and Castelló (2017) used PCA to solve the collinearity problem among the independent variables in multivariate analysis for clinical studies, where instead of including correlated independent variable; two uncorrelated principal components were used. The algorithm that does not utilise all the conceivable features is the decision tree algorithm which utilises significant features and overlooks insignificant ones (Witten et al., 2016; Witten, 1999). In the present paper, we have not eliminated the data instances having missing data and the maximum information are retained due to multiple imputation.

3 Materials and methods

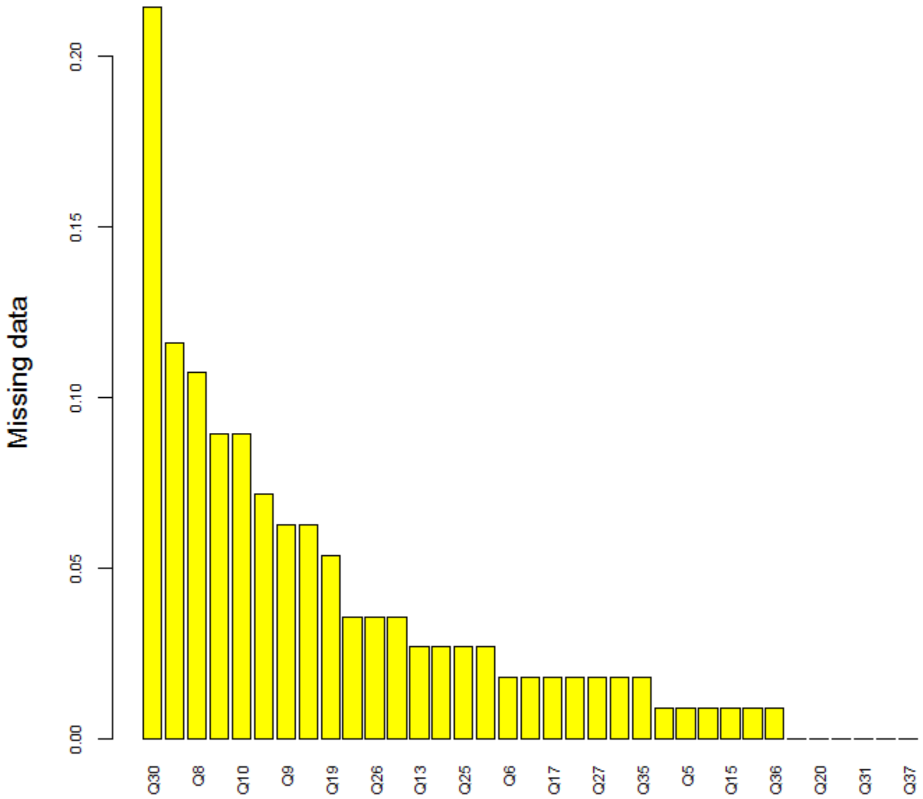
In this research, we used different factors which are responsible for obese and overweight and ranked the responses. Also, identifies the normal or abnormal body weight without knowing the age and BMI, and only with the help of food habits. Hence, this type of research gains importance. In this research, a perceptions-only measure is used to collect data from women of different age groups, especially working women and teenage girls. A five-point scale (5 indicating strongly agree and 1 indicating strongly disagree) was used in preference to a seven-point scale to increase the sensitivity of the measure. In this study, normality and abnormality of the body weight were measured with a self-administered questionnaire. The data of fifty women belonging to different age were obtained from the staff and students of a women's educational institution. The data is obtained by the classical method of collecting input by means of a questionnaire. The questionnaire consists of 35 questions in total, which is the combination of biological, psychological, types of food and many other different factors of food habits.

Collected data were analysed with the help of software tool R. Statistical techniques like PCA were used to reduce the factors and trained using the classifier algorithms. Some of the biological questions like their age, weight, and height were excluded for training and testing. The questions related to psychological behaviour and food eating habits alone are taken into consideration.

3.1 Missing data imputation

A survey cum experimental methodology is used to collect the data. The first obstacle in predictive modelling is considered to be missing values. Figure 1 shows the visual representation of the missing values. The histogram depicts the influence of missing value in the variable. There are 21% missing values in question no. 30 and 19% missing value in question no. 1 and so on. In this paper, the missing data are missing at random (MAR), which means that the probability that a value is missing depends only on observed value and can be predicted using them. It imputes data on a variable by variable basis by specifying an imputation model per variable. Generating more than one imputation compared to a single imputation looks after uncertainty in missing values. If X_1 has missing values, then it will be regressed on other variable X_2 to X_k. The missing values in X_1 will be then replaced by predictive values obtained.

Figure 1 Visual representation of missing data (see online version for colours)



To predict missing values, linear regression and logistic regression are used for categorical missing values. More than one imputations do not attempt to estimate every missing value via simulated values however alternatively to symbolise a random sample of the missing values (Yuan, 2010). In this paper, predictive mean matching (PMM) method is used to predict numeric variable. PMM produces imputed values that are much more like real values. If the original variable is skewed, the imputed values will also be skewed. If the original variable is between zero and hundred, the imputed values will also be between zero and hundred. And if the real values are discrete (like a number of children), the imputed values will also be discrete. That is because the imputed values are actual values which might be borrowed from individuals with real data.

3.2 The PCA approach

1 standardise the data

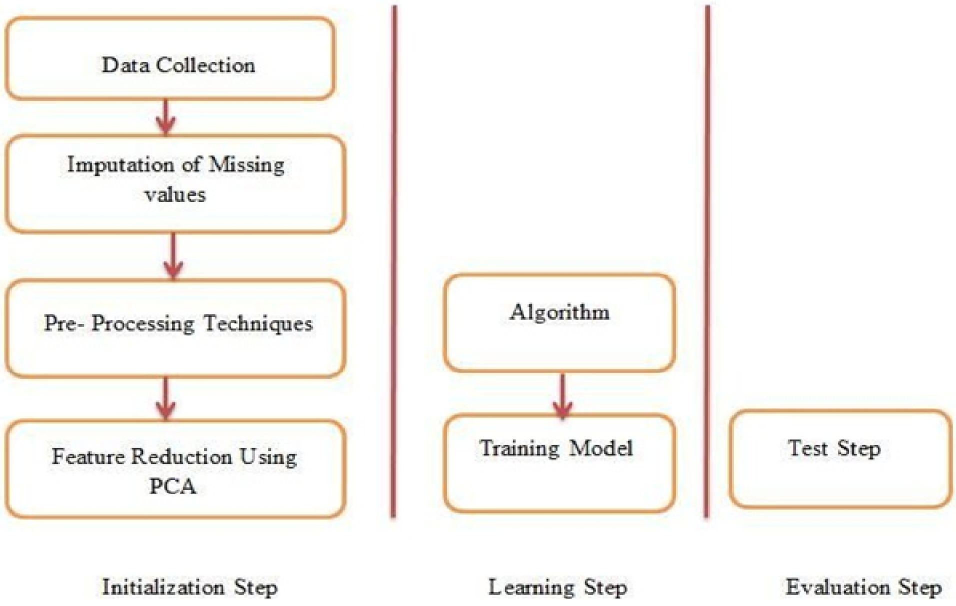
2 obtain the eigenvectors and eigenvalues from the covariance matrix or correlation matrix

3 sort eigenvalues in descending order and choose the k eigenvectors that correspond to the k largest eigenvalues where k is the number of dimensions of the new feature subspace (k <= d)

4 construct the projection matrix W from the chosen k eigenvectors

5 transform the original dataset X through W to acquire a k-dimensional feature subspace Y.

Figure 2 Block diagram of the proposed model (see online version for colours)



3.3 The prediction hypothesis

The outcome of training a classifier is a hypothesis that can be used for predictions. By training NB, SVM, DT, KNN and LR, we obtained three hypotheses for predicting the outcome of a match. In order to assess the effectiveness of these hypotheses, we performed evaluation task as described below. To evaluate the performance of classifiers, the following steps have been performed:

1 Imputation of missing value using PMM method.

2 Standardise numerical data (e.g., mean of 0 and standard deviation of 1) using the scale and centre options.

3 Normalise numeric data (e.g., to a range of 0–1) using the range option.

4 Using PCA for factor reduction.

5 Training the normalised data along with the selected factors using classifier algorithms.

6 Calculating and evaluating the accuracy. Accuracy percentage can be defined as follows, accuracy percentage = (TC/N) * 100, where N is the total number of test cases, TC is the total number of subjects correctly classified.

7 Calculate the AUC.

3.4 Classifier training

A brief introduction to classifier algorithms NB, SVM, decision tree, KNN and backpropagation neural network model is given below.

3.4.1 Naive Bayes algorithm

The naive Bayes classifier greatly simplifies learning by assuming that features are independent to given class (Rish, 2001). Class conditional independence in naive Bayes classifiers assures that value of one attribute is independent of another attribute. Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, naive Bayes is known to outperform even highly sophisticated classification methods. Bayes theorem provides a way of calculating posterior probability P(c | x) from P(c), P(x) and P(x | c). Look at the equation below:

$$P(c \mid x) = P(x \mid c)\,P(c)/P(x) \tag{1}$$

Above

- P(c | x) is the posterior probability of class (c, target) given predictor (x, attributes).
- P(c) is the prior probability of class.
- P(x | c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor.

3.4.2 Support vector machine

Support vector machine is a suitable method for classification of linear and nonlinear data (Tong and Koller, 2001). Support vector machine is an algorithm that converts training data into a higher dimension using nonlinear mapping. It searches for the linear and optimal separating hyper plane within this new dimension. Data from two classes are separated by a hyper plane. The two components used by SVM to find the hyper plane are support vectors and margins. SVM are less vulnerable to over fitting than other classifiers. The application of SVM includes handwritten digit recognition, object recognition, etc.

3.4.3 Decision tree

Decision trees are one of the most popular and powerful approaches in data mining. Decision tree are used mostly in decision theory and statistics. A decision tree is a tree like structure where each internal node represents test on an attribute, each branch represents the results of the test and each leaf represents the class label. The benefits of a decision tree in data mining:

1 it can able to handle the variety of input data such as nominal, numeric and textual

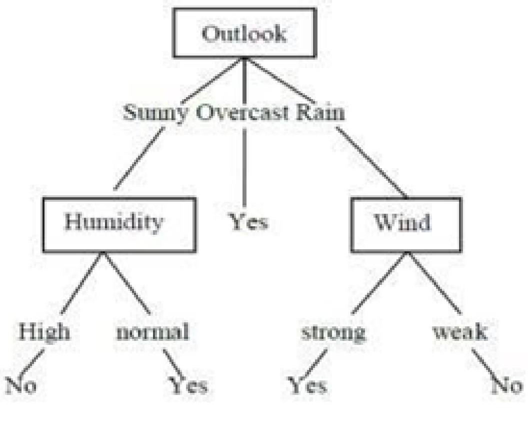
2 it processes the dataset that contains the errors and missing values

3 it is available in vary packages of data mining and number of platform (Li and Zhang, 2010).

Decision tree classifies the new data efficiently by tracing the path from the root to leaf node.

The following example illustrates working of decision tree algorithm (Chourasia, 2013).

Figure 3 Decision tree



3.4.4 K-nearest neighbourhood

K-nearest neighbour method is a lazy learner algorithm, which is widely used in the area of pattern recognition. In this algorithm, the training tuples are represented by n attributes. This algorithm compares the given test tuple with training tuple that are similar to it. For an unknown sample to be classified, the pattern space of the k training tuples will be searched that are closest to the unknown tuple. These k training samples are the k nearest neighbours of the unknown sample. To define the closeness, compute the Euclidean distance of each object in the dataset from each of the centroids (Peterson, 2009).

3.4.5 Back propagation neural network

Back propagation is a neural network algorithm. In neural networks, input and output units are connected to each other. Each unit has a weight associated with it. In the learning phase, the neural networks learn by adjusting the weights, either by reward or penalty, to predict the class label of the unknown tuple. The time it takes to train the data set is quite long compared to other classification algorithm. Parallelisation techniques can be adopted to speed up the process. This algorithm is very much suitable for continuous data. The neural network can be used when we have little knowledge on the relationship between the attributes and classes (Li et al., 2012).

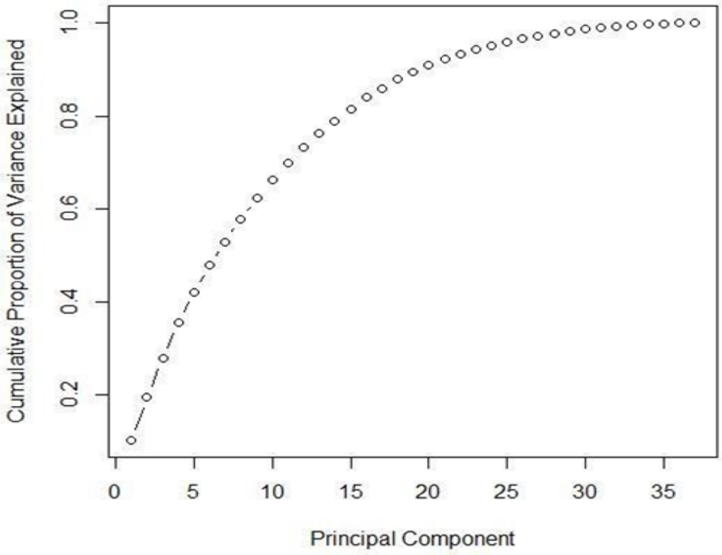
4 Results

The dataset is obtained using survey cum questionnaire method through the expert and document review through 37 questions. The result of the first step is the imputation datasets. In this paper, multiple imputation method is adopted. Five different datasets are generated after imputation. These datasets are pre-processed using centre, range, scale, centre and scale method. After normalisation, to interpret the data in a more meaningful form, it is therefore necessary to reduce the number of variables to a few, interpretable linear combinations of the data.

4.1 Results of PCA

Each linear combination will correspond to a principal element. The filtering technique PCA was used to reduce the dimension obtained from the third step by combining the related variables to be the new variables called factors. In this dataset, it reduces the dimensions of the data from 37 to 29 variables. Figure 4 shows the cumulative scree plot after applying PCA.

Figure 4 Cumulative scree plot



The first 29 principal components explain 96% of the variation. This is an acceptably large percentage. Because of standardisation, all principal components will have mean 0. The standard deviation is also given for each of the components and these will be the square of the eigenvalue. More important for our current purposes are the correlations between the principal components and the original values.

4.2 Classifier evaluation

The experimentation of this work was performed using R tool. R has been chosen due to several reasons: it has built-in-state of art feature selection, classification and evaluation methods. After multiple imputations using PMM method, five different dataset has been generated. These dataset are pre-processed using different pre-processing methods like centre, scale, range, centre and scale. The dataset is transformed into factors, i.e., principal components. This dataset generated PC1 to PC37. The principal components PC1 to PC 29 has been chosen after the transformation. For training and evaluating the classifiers, these datasets are used along with various classification algorithms like NB, SVM, decision tree, KNN, and NNET. The results are tabulated in Tables 2–5.

Table 2 Classifier performance using PCA after imputation and pre-processed with 'scale'

Imputed dataset	*Classifier*	*Accuracy*	*AUC*	*Time elapsed*
IMPUTED_DATASET_1	NB	41%	56%	0.1326
	SVM	44%	54%	0.0498
	Decision tree	47%	52%	0.047
	KNN	61%	61%	1.0766
	BPN-NNET	54%	55%	6.1878
IMPUTED_DATASET_2	NB	68%	69%	0.1319
	SVM	59%	62%	0.0467
	Decision tree	62%	65%	0.0463
	KNN	61%	56%	1.0754
	BPN-NNET	63%	62%	6.1352
IMPUTED_DATASET_3	NB	65%	65%	0.1456
	SVM	53%	53%	0.0549
	Decision tree	53%	53%	0.048
	KNN	70%	70%	1.1668
	BPN-NNET	60%	60%	6.6999
IMPUTED_DATASET_4	NB	70%	72%	0.1411
	SVM	65%	67%	0.0497
	Decision tree	65%	63%	0.0517
	KNN	65%	64%	1.1294
	BPN-NNET	61%	63%	6.5273
IMPUTED_DATASET_5	NB	50%	50%	0.1433
	SVM	59%	59%	0.0504
	Decision tree	59%	58%	0.046
	KNN	66%	67%	1.1544
	BPN-NNET	59%	59%	6.4505

Table 3 Classifier performance using PCA after imputation and pre-processed with ‘centre’

Imputed dataset	*Classifier*	*Accuracy*	*AUC*	*Time elapsed*
IMPUTED_DATASET_1	NB	50%	52%	0.166
	SVM	65%	65%	0.0469
	Decision tree	50%	53%	0.0442
	KNN	63%	63%	1.0368
	BPN-NNET	73%	74%	5.8211
IMPUTED_DATASET_2	NB	68%	67%	0.1425
	SVM	74%	74%	0.0493
	Decision tree	44%	57%	0.0504
	KNN	67%	66%	1.1257
	BPN-NNET	69%	70%	6.524
IMPUTED_DATASET_3	NB	65%	66%	0.1357
	SVM	59%	60%	0.0499
	Decision tree	35%	63%	0.0519
	KNN	70%	70%	1.0927
	BPN-NNET	58%	58%	6.1808
IMPUTED_DATASET_4	NB	65%	63%	0.1404
	SVM	76%	76%	0.0507
	Decision tree	65%	58%	0.0507
	KNN	67%	67%	1.134
	BPN-NNET	75%	75%	6.477
IMPUTED_DATASET_5	NB	65%	64%	0.1461
	SVM	65%	65%	0.0518
	Decision tree	65%	66%	0.0482
	KNN	77%	77%	1.1685
	BPN-NNET	71%	70%	6.5245

Table 4 Classifier performance using PCA after imputation and pre-processed with ‘range’

Imputed dataset	*Classifier*	*Accuracy*	*AUC*	*Time elapsed*
IMPUTED_DATASET_1	NB	59%	58%	0.1389
	SVM	56%	56%	0.0478
	Decision tree	38%	62%	0.0473
	KNN	51%	55%	1.0846
	BPN-NNET	60%	60%	6.0882
IMPUTED_DATASET_2	NB	56%	55%	0.1391
	SVM	53%	51%	0.0495
	Decision tree	47%	54%	0.0526
	KNN	45%	54%	1.1075
	BPN-NNET	57%	57%	6.1637

Table 4 Classifier performance using PCA after imputation and pre-processed with 'range' (continued)

Imputed dataset	*Classifier*	*Accuracy*	*AUC*	*Time elapsed*
IMPUTED_DATASET_3	NB	56%	60%	0.149
	SVM	55%	55%	0.0588
	Decision tree	50%	50%	0.054
	KNN	64%	64%	1.1938
	BPN-NNET	63%	63%	6.6065
IMPUTED_DATASET_4	NB	65%	65%	0.1406
	SVM	68%	68%	0.0521
	Decision tree	56%	56%	0.0493
	KNN	64%	65%	1.1353
	BPN-NNET	69%	69%	6.3515
IMPUTED_DATASET_5	NB	59%	61%	0.1431
	SVM	59%	61%	0.0541
	Decision tree	59%	58%	0.0472
	KNN	58%	58%	1.1797
	BPN-NNET	70%	71%	6.5004

Table 5 Classifier performance using PCA after imputation and pre-processed with 'centre and scale'

Imputed dataset	*Classifier*	*Accuracy*	*AUC*	*Time elapsed*
IMPUTED_DATASET_1	NB	62%	62%	0.1338
	SVM	71%	70%	0.0469
	Decision tree	74%	73%	0.0456
	KNN	60%	61%	1.0597
	BPN-NNET	63%	63%	5.8788
IMPUTED_DATASET_2	NB	62%	62%	0.1288
	SVM	59%	59%	0.0489
	Decision tree	53%	54%	0.0456
	KNN	45%	56%	1.0741
	BPN-NNET	66%	66%	5.9304
IMPUTED_DATASET_3	NB	65%	65%	0.1415
	SVM	65%	65%	0.0507
	Decision tree	56%	56%	0.0484
	KNN	58%	58%	1.1312
	BPN-NNET	69%	69%	6.2898

Table 5 Classifier performance using PCA after imputation and pre-processed with 'centre and scale' (continued)

Imputed dataset	*Classifier*	*Accuracy*	*AUC*	*Time elapsed*
IMPUTED_DATASET_4	NB	62%	63%	0.1461
	SVM	59%	63%	0.0506
	Decision tree	53%	53%	0.0483
	KNN	64%	62%	1.1532
	BPN-NNET	61%	63%	6.3367
IMPUTED_DATASET_5	NB	47%	50%	0.139
	SVM	65%	67%	0.0504
	Decision tree	68%	68%	0.0479
	KNN	59%	59%	1.1209
	BPN-NNET	60%	62%	6.4684

5 Discussion

The present study shows that the accuracy shown in Table 3 are found to be better than the accuracies of other classifier when imputed with different pre-processing techniques using PCA. In Table 3, in 'centre', we calculate the average value of each variable and then subtract it from the data. This implies that each column will be transformed in such a way that the resulting variable will have a zero mean. For the five imputed datasets, the highest obtained accuracy are for BPN, SVM, KNN, SVM and KNN respectively with 73%, 74%, 70%, 76% and 77%.

The data pre-processing method along with the classifier which gives highest accuracy has been tabulated in Table 6.

Table 6 Classification accuracies of proposed model

Dataset	*Name of the classifier*	*Pre-processing technique*	*Accuracy*
1	Back propagation neural network	Centre	73%
2	Support vector machine	Centre	74%
3	K-nearest neighbour	Centre, scale	70%
4	Support vector machine	Centre	76%
5	K-nearest neighbour	Centre	77%

6 Conclusions

The aim of this research is to carry out the feature selection using PCA to determine the normal and abnormal body weight of women. This work adopts a straightforward methodology. The dataset consists of some missing values. It is imputed by multiple imputations. The imputed data are pre-processed using different methods. The feature reduction on normalised data is then made by PCA. Thirty seven features are reduced to 29. Then the model is generated using different classifier with the different pre-processing method. Our results show that centering the data to its mean gives a satisfactory accuracy percentage. On the other hand, this data needs the highest accuracy percentage. Therefore, in future work, the other algorithms used for feature selection will be studied; for example, the other feature selection method like filter methods, wrapper methods, and embedded methods can be used.

References

Abolmakarem, S., Abdi, F. and Khalili-Damghani, K. (2016) 'Insurance customer segmentation using clustering approach', *International Journal of Knowledge Engineering and Data Mining*, Vol. 4, No. 1, pp.18–39.

Agarap, A.F.M. (2018) 'On breast cancer detection: an application of machine learning algorithms on the wisconsin diagnostic dataset', in *Proceedings of the 2nd International Conference on Machine Learning and Soft Computing*, ACM, February, pp.5–9.

Ahmad, A., Mustapha, A., Zahadi, E.D., Masah, N. and Yahaya, N.Y. (2011) 'Comparison between neural networks against decision tree in improving prediction accuracy for diabetes mellitus', in *Digital Information Processing and Communications*, pp.537–545, Springer, Berlin, Heidelberg.

Amatul, Z., Asmawaty, T., Kadir, A. and Aznan, M.A.M. (2013) 'A comparative study on the pre-processing and mining of Pima Indian diabetes dataset', in *3rd International Conference on Software Engineering & Computer Systems (ICSECS – 2013)*, Universiti Malaysia Pahang, pp.1–10, 20–22 August.

Chen, W., Chen, S., Zhang, H. and Wu, T. (2017) 'A hybrid prediction model for type 2 diabetes using K-means and decision tree', in *2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, IEEE, November, pp.386–390.

Chourasia, S. (2013) 'Survey paper on improved methods of ID3 decision tree classification', *International Journal of Scientific and Research Publications*, Vol. 3, No. 12, pp.1–4.

Guyon, I. and Elisseeff, A. (2003) 'An introduction to variable and feature selection', *Journal of Machine Learning Research*, March, Vol. 3, pp.1157–1182.

Hall, M.A. (1999) *Correlation-Based Feature Selection for Machine Learning*, University of Waikato, Hamilton.

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I. and Chouvarda, I. (2017) Machine learning and data mining methods in diabetes research', *Computational and Structural Biotechnology Journal*, Vol. 15, No. 1, pp.104–116.

Li, J. et al. (2012) 'Brief introduction of back propagation (BP) neural network algorithm and its improvement', *Advances in Computer Science and Information Engineering*, pp.553–558.

Li, L. and Zhang, X. (2010) 'Study of data mining algorithm based on decision tree', *International Conference on Computer Design and Applications (ICCDA)*, IEEE, Vol. 1, pp.V1–155.

Liu, H. and Motoda, H. (2012) *Feature Selection for Knowledge Discovery and Data Mining*, Vol. 454, Springer Science & Business Media, New York.

Liu, X., Dai, Y., Zhang, Y., Yuan, Q. and Zhao, L. (2017) 'A preprocessing method of AdaBoost for mislabeled data classification', in *2017 29th Chinese Control and Decision Conference (CCDC)*, IEEE, May, pp.2738–2742.

Manek, S., Vijay, S. and Kamthania, D. (2016) 'Educational data mining – a case study', *International Journal of Information and Decision Sciences*, Vol. 8, No. 2, pp.187–201.

Nayak, A.S., Kanive, A.P., Chandavekar, N. and Balasubramani, R. (2016) 'Survey on pre-processing techniques for text mining', *International Journal of Engineering and Computer Science*, June, Vol. 5, No. 6, pp.16875–16879.

Patil, B.M., Joshi, R.C. and Toshniwal, D. (2010) 'Hybrid prediction model for type-2 diabetic patients', *Expert Systems with Applications*, Vol. 37, No. 12, pp.8102–8108.

Peña, J.M., Lozano, J.A., Larrañaga, P. and Inza, I. (2001) 'Dimensionality reduction in unsupervised learning of conditional Gaussian networks', *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 1 June, Vol. 6, pp.590–603.

Peterson, L.E. (2009) 'K-nearest neighbour', *Scholarpedia*, Vol. 4, No. 2, p.1883.

Rish, I. (2001) 'An empirical study of the naive Bayes classifier', in *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, IBM, Vol. 3, No. 22.

Suranovic, S., Goldfarb, R.S. and Leonard, T.C. (2003) *An Economic Analysis of Weight Change, Overeating and Dieting*, Unpublished Working Paper.

Tong, S. and Koller, D. (2001) 'Support vector machine active learning with applications to text classification', *Journal of Machine Learning Research*, November, Vol. 2, pp.45–66.

Uma, K. and Hanumanthappa, M. (2017) 'Data collection methods and data pre-processing techniques for healthcare data using data mining', *International Journal of Scientific &Engineering Research*, Vol. 8, No. 6, pp.1131–1136.

Witten, I.H., Frank, E., Hall, M.A. and Pal, C.J. (2016) *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, USA.

Wu, H., Yang, S., Huang, Z., He, J. and Wang, X. (2018) 'Type 2 diabetes mellitus prediction model based on data mining', *Informatics in Medicine Unlocked*, January, Vol. 10, pp.100–107.

Yu, L. and 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.

Yuan, Y.C. (2010) *Multiple Imputations for Missing Data: Concepts and New Development (Version 9.0)*, Vol. 49, pp.1–11, SAS Institute Inc., Rockville, MD.

Zhang, Z. and Castelló, A. (2017) 'Principal components analysis in clinical studies', *Annals of Translational Medicine*, Vol. 5, No. 17, p.351.