**Journal of Biomedical Engineering**
ISSN: 1001-5515

MANUSCRIPT
Vol. 40 No. 3 (2023)

# BAYESIAN LOGISTIC REGRESSION WITH ROBUST PREDICTORS FOR GESTATIONAL DIABETES PREDICTION

**David Kwamena Mensah[1], Etornam Kwame Kunu[2], Francis Eyiah-Bediako[3], Samuel Assabil[4] and Richard Okyere[5]**

[1,2,3,4] Department of Statistics, University of Cape Coast, Cape Coast, Ghana

[5] Ghana Insurance College, Accra, Ghana

*Corresponding author: **dmensah@ucc.edu.gh**

**Abstract**

The presence of outlying observations within the predictor space of the dataset for binary logistic regression can impact significantly the predictive performance of the developed classifier for public health problems. In this regard, this paper considers improving the predictive performance of the binary logistic classifier for Gestational Diabetes Mellitus prediction using alternative predictors termed robust predictors. These predictors are based on the computation of non-central moments of probability density functions of the original predictors. With this treatment, the relative importance of predictor-specific observations is easily assessed with outlying observations handled automatically and utilized for model fitting without being deleted. This way, the predictors become robust to extreme observations, and predictor-specific autocorrelations, allowing easy extension of binary logistic classifiers to public health problems for which outlying observations are inevitable. Appropriate Binary logistic regression models using the alternative predictors were developed within both the Classical and Bayesian paradigms with their utility illustrated with both simulated datasets and real gestational diabetes mellitus datasets, in comparison with existing current proposals.

**Keywords**: Gestational diabetes mellitus, Logistic regression, Bayesian inference, Alternative predictors, Kernel density estimation

## 1 Introduction

Diabetes is one of the key public health problems worldwide due to its contribution to cause of death. This disease is rated the ninth leading cause of death globally [WHO, 2020]. Also, the International Diabetes Federation (IDF) reports that about 463 million cases of diabetes occurred worldwide [IDF, 2020]. Fortunately, or unfortunately, approximately 50.09% of these cases remained undiagnosed [IDF, 2020]. In the light of this, prompt diagnosis and subsequent treatment are regarded as the key measures and interventions needed to mitigate complications that lead to a substantial decrease in patients' quality of life and even death which are usually preventable through timely detection and identification of risk factors [Bantie et al., 2019, IDF, 2020]. As a result, diabetes modeling and prediction have piqued the interest of researchers within the statistical community.

In recent times, cutting-edge techniques for timely identification and prediction of diabetes mellitus has been explored based on the exploration of existing machine learning (ML) algorithms



DOI: 10.105515/JBE.40.3.41

to gain useful insights into accessible clinical diabetes data. Notable of these techniques are the Binary logistic regression, Support vector machine (SVM), Random Forest Decision tree, K-Nearest Neighbors and Naïve Bayes models [Kavakiotis et al., 2017, Ahuja et al., 2019, Mujumdar and Vaidehi, 2019].

It has been established that the predictive performance of binary logistic regression model can be impaired by the presence of extreme observation(s) as well as noise in a binary dataset [Gelman, 2007]. Motivated by the observation made by Gelman [2007], the literature has registered some exploration of the integration or joint-use of existing standard statistical methods such as the K-Means clustering, and Principal Component Analysis (PCA) in the ML algorithms.

Iyer et al. [2015] proposed decision tree and Naïve Bayes models for diabetes prediction, in which result yielded an efficient model with some level reduction in error rate. Jhaldiyal and Mishra [2014] considered the integration of principal component analysis (PCA) into a support vector machine (SVM) algorithm for patient diabetes disease state classification. Their approach led to the realization that PCA with SVM performs well for diabetes mellitus prediction, leading to the attainment of about 93.66% predictive accuracy. Wu et al. [2018] incorporated the K-means into the logistic regression classier to model and predict diabetes. This approach also exhibited an improvement in the predictive accuracy with about 95.42% accuracy.

Recently, Zhu et al. [2019] considered the use of both the PCA and K-Means clustering within the binary logistic regression framework. In particular, the dataset was reduced by applying the PCA technique followed by K-Means clustering. The resulting clustered dataset was used to fit a binary logistic regression. This led to an efficient binary logistic regression model for the early prediction of diabetes using the Pima Indian Diabetes dataset. Their approach suggested an improved logistic regression model for predicting diabetes, with 2.02% improvement over that of [Wu et al., 2018]. This suggests that the integration of the PCA and K-Means clustering technique improved the classification accuracy of logistic regression for the Pima Indian Diabetes dataset.

The joint use of PCA and K-Means in logistic regression can be viewed as the current state-of-the-art approach for improving the performance of the logistic classifier for diabetes modeling. However, though these techniques are appealing, there exist some drawbacks. First, they are susceptible to outliers and, as such, may still produce incorrect classification in the presence of data points that deviate from the expected range of values. This is because both K-Means and PCA are based on a measure of center (mean), which is not robust. For the PCA, the application of the above center is seen in the computation of covariance profiles. Thus, their use in logistic regression, regardless of the modeling framework, still inherits these drawbacks. Therefore, though the incorporation of the PCA and K-Means into the logistic regression improved the model classification accuracy in the studies of Zhu et al. [2019], they do not control for outliers. As a result, the improved accuracy observed may be due to the richness of the data or that the data contains no outliers. For instance, the K-Means are limited to some extent by the presence of outliers which leads to misclassification due to the outliers' ability to drag the proposed centroid.


**DOI**: 10.105515/JBE.40.3.41

Second, the joint use of PCA and K-means in logistic regression may introduce extra computational expenditure. This is because computational challenges resulting from the use of a large dataset cannot be ruled out. Third, by principle, PCA is a data reduction technique that is meant to reduce the dimension of a large dataset for computational savings. Unfortunately, the reduced dataset offered by the application of PCA can still possess outliers if robust centers are not considered. In summary, PCA is not meant to control outliers; thus, outliers may still exist among the reduced dimensions. In particular, the PCA depends on the mean, which is sensitive to outliers for the computation of covariance matrices, thereby posing a drawback. Thus, reducing the data without controlling outliers might lead to a vital loss of information. Moreover, the reduced data might come with outlying information due to their inability to control them.

In general, it is natural for every statistical data to have some outliers present regardless of the type and mode of collection. However, these outliers may constitute a valuable source of information or otherwise. Thus, if not properly handled can distort many measures in the data analysis and statistical modeling, thus disproportionate the estimated values of model parameters. Nevertheless, varied statistical techniques for modeling statistical data in the presence of outliers are challenged in terms of direct application. It is important to note that these observations can occur within the predictor or response space. That is a response can be an outlier as well as a predictor. In most cases, outliers are controlled within the response space but not the predictor space. Actually, deletion of an outlier may lead to loss of information or reduced information content of the data at hand. On the other hand, its inclusion in model specifications can pose challenges.

Another appealing line of improvement on the predictive performance of binary logistic regression model in the Bayesian framework has been considered by [Asanya et al., 2021]. Asanya et al. [2021] proposed treatment for outliers in binary logistic regression, adopting the Bayesian approach with student *t* prior distributions considered robust for parameters in the case of small sample size. The authors claim that the student *t* distribution serves as a shock absorber to the outliers and other random fluctuations; thus providing robustness to the model. Their work was motivated by the work of Gelman [2007]. It is important to note that their application was in the direction of Immunotherapy where a dataset on wart treatment was used. Though their approach is novel, the treatment for outliers is seen in the response and parameter spaces but not within predictor space. The presence of outliers within the predictor can still affect the performance of the binary logistic regression model. This paper focuses on the development of computational methods that offer automatic outlier control and data reduction for enhanced predictive disease modeling. In particular, in this paper, robust priors similar to those used by Asanya et al. [2021] and robust predictors are considered within the Bayesian framework. We argue that statistical methods that can ensure automatic control for outliers both in the predictor and response spaces have the potential for substantial improvement in predictive performance. The rest of the paper is structured as follows. Section 2, formally introduces the binary logistic regression models based on the original data predictors as generative models for gestational diabetes mellitus data. Section 3 gives a brief exposition on the alternative predictors for binary logistic regression with extraction


DOI: 10.105515/JBE.40.3.41

schemes based on the first non-central moment statistic formulated. Section 4 treats the incorporation of the alternative predictors into the proposed binary logistic regression models. Section 5 outlines both classical and Bayesian inference methods for fitting and inference for the developed models. Appropriate performance measures for assessing the binary logistic regression models are discussed in Section 6. Section 7 presents the description of the Pima Indian Diabetes dataset, as well as data preprocessing methods. Section 8 gives a brief on implementation of the methods. Some examples in simulation and real data applications are illustrated in Section 9 and Section 10 concludes the paper.

## 2 Methods

Let $y=\left[y_1,y_2,\ldots,y_m\right]$, $y\in(0,1), i=1,\ldots,m$ with $y_i=1$ denoting the *i*th patient is diabetic and $y_i=0$ denoting *i*th patient non-diabetic. Also, let $X_p=\left[X_1,X_2,\ldots,X_p\right]$, $X_j=\left[x_{1j},x_{2j},\ldots,x_{pj}\right]$, $j=1,\ldots,p$ denotes a *p*-dimensional covariate (predictor) vector predicting *y*. Write $X=\left[1,X_p\right]$, where 1 denotes an $m$ column vector of 1s for an $m\times(p+1)$ design matrix, where we have written $a'$ as the transpose of the vector $a$. Write the success and failure probabilities as $p(y_i=1)=\vartheta_i$, and $p(y_i=0)=1-\vartheta_i$ respectively. Then, the generative model for $y_i$ is assumed to follow the Bernoulli probability model

$$p(y_i\mid\vartheta_i)=\vartheta_i^{y_i}(1-\vartheta_i)^{1-y_i},0<\vartheta_i<1. \tag{1}$$

We further model the set of predictors, $X_p$ via a logit link function

$$\log\left(\frac{\vartheta_i}{1-\vartheta_i}\right)=X_i'\beta \tag{2}$$

where $X_i=\left[1,x_{i1},x_{i2},\ldots,x_{ip}\right]$, $\vartheta_i=\frac{e^{X_i'\beta}}{1+e^{X_i'\beta}}, 1-\vartheta_i=\left(1+e^{X_i'\beta}\right)^{-1}$, $\beta=\left[\beta_0,\beta_1,\ldots,\beta_p\right]$ is a set of regression coefficients of medical attribute predictors.

Accordingly, the sampling distribution of the data can be expressed as

$$p(y\mid X,\beta)=\prod_{i=1}^{m}\left[\frac{e^{X_i'\beta}}{1+e^{X_i'\beta}}\right]\left[\left(1+e^{X_i'\beta}\right)^{-1}\right]^{1-y_i} \tag{3}$$

Suppose there exist outliers in the predictor space (i.e among the $X_p$s). Then, they need to be carefully handled but not deleted. We introduce a simple approach for handling them in the next Section

## 3 Robust predictors for logistic regression

Recall the design matrix $X$ with defining covariate vector $X_p$. There is a possibility for each or some of these covariates to possess observations acting differently from the overall pattern underlying the predictor data. Deletion of such observation may lead to loss of information and


DOI: 10.105515/JBE.40.3.41

their inclusion may also influence the inference making process in some way. In fact, the outlying predictor observation if there exists constitute a vital information in terms of the defining features of the predictor data. Thus, there is the need to handle them in a controlled manner in model specification such that their contributions to the underlying structures in the data can be utilized. We consider the use of alternative covariates derived from the computation of non-central moments of random variables (probability density functions). The basic idea is to map the original predictors unto the predictor specific probability density functions space underlying the predictor data and use features based on the first non-central moments. In effect, by assessing the contributions of the predictors with the help of the density functions, robust covariates can be derived in place of the original. This leads to a principled probabilistic approach for determining vital covariates as well as controlling for outlying ones without deleting them. In what follows, we illustrate how the alternative covariates are derived. For each continuous covariate say $X_j$ where $j = 1,2, \ldots, p,$ we compute a statistic based on a non-central moment,

$$C\left(X_j\right) = x_j f\left(x_j\right).$$

The motivation for the above statistics is based on interesting features of moments of random variables or probability distributions.

Consider the *kth* non-central moments of $X_j$.

$$E\left[X_j^k\right] = \int_{-\infty}^{\infty} x_j^k f\left(x_j\right) dx_j \qquad (4)$$

The integrand provides a natural way to access the contribution of each observation $x_{ij}$ towards the common measure of center. With this, it is easily observed that observation with high contribution to a common center will be clustered around the center. Also, a contribution deviating appreciably from the underlying data structure will be located at the tails of the pdf. Thus, $f\left(x_j\right)$ serves as natural predictor observation-specific weight, weighting appropriately the observations such that extreme ones are lowly weighted limiting their impact on the common center. Based on the above observations, the statistic $C\left(X_j\right)$ can be seen as appropriate to serve as an alternative covariate for logistic modelling. Thus, instead of using the $X_j's$, we use the $C\left(X_j\right)s$ in modelling.

The new covariate data structure will be of the form indicated in Table 1


DOI: 10.105515/JBE.40.3.41

| Obs. | $X_1$ | $X_2$ | $\cdots$ | $X_p$ | | $c(X_1)$ | $c(X_2)$ | $\cdots$ | $c(X_p)$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | $x_{11}$ | $x_{21}$ | $\cdots$ | $x_{p1}$ | | $c(x_{11})$ | $c(x_{21})$ | $\cdots$ | $c(x_{p1})$ |
| 2 | $x_{12}$ | $x_{22}$ | $\cdots$ | $x_{p2}$ | | $c(x_{12})$ | $c(x_{22})$ | $\cdots$ | $c(x_{p2})$ |
| 3 | $x_{13}$ | $x_{23}$ | $\cdots$ | $x_{p3}$ | $\overrightarrow{Transformed}$ | $c(x_{13})$ | $c(x_{23})$ | $\cdots$ | $c(x_{p3})$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| m | $x_{1m}$ | $x_{2m}$ | $\cdots$ | $x_{pm}$ | | $c(x_{1m})$ | $c(x_{2m})$ | $\cdots$ | $c(x_{pm})$ |

Furthermore, the corresponding design matrix can be expressed as

$$C(X)=\left[1,\ C\left(X_1\right),C\left(X_2\right),\ldots,C\left(X_p\right)\right] \tag{5}$$

where $C\left(X_j\right)=\left[C\left(X_{1j}\right),C\left(X_{2j}\right),\ldots,C\left(X_{mj}\right)\right]'$.

**4 Application of robust covariates in logistic regression**

Now, applying the robust covariates (5), the model (2) becomes

$$\log\left(\frac{\vartheta_i}{1-\vartheta_i}\right)=C\left(X_i\right)\beta \tag{6}$$

This leads us to the likelihood function defined as

$$p\left(C\left(X\right),y\mid\beta\right)=\prod_{i=1}^{m}\left[\frac{e^{C(X_i)\beta}}{1+e^{C(X_i)\beta}}\right]^{y_i}\left[\left(1+e^{C(X_i)\beta}\right)^{-1}\right]^{1-y_i} \tag{7}$$

Consequently, the log-likelihood function of based on the robust covariates is

$$\prod_{i=1}^{m}y_i\log\left[\frac{e^{C(X_i)\beta}}{1+e^{C(X_i)\beta}}\right]+\left(1-y_i\right)\log\left[\left(1+e^{C(X_i)\beta}\right)^{-1}\right] \tag{8}$$

**5 Inference for logistic regression model**

**5.1 Classical inference for Binary logistic regression**

In practice, deriving the maximum likelihood estimates (MLE) for logistic regression model is analytically complicated, hence, numerical methods are applied to compute the global maximizer $\beta_j$ in (8). In this study, we derived the MLE via the fisher scoring method. This is an iterative technique for deriving solutions to likelihood equations [see, Agresti, 2015]. In particular, it replaces the Hessian matrix (observed information) in Newton-Raphson method with the expected information matrix. The fisher scoring method is implemented in most packages in R. The application here considered the *glm* function in R.

**5.2 Bayesian Approach for Logistic Regression**


DOI: 10.105515/JBE.40.3.41

Under the Bayesian framework, the estimation of model parameters is done via a posterior distribution $p(\beta \mid C(X), y)$, given some prior knowledge, $p(\beta)$. Thus, given the prior knowledge, $p(\beta)$ for $\beta$, likelihood function, $p(y \mid \beta)$, and a data vector *y*, the Bayesian inference about the parameter $\beta$ is based on the posterior distribution given as

$$p(\beta \mid C(X), y) = \frac{p(y \mid \beta)\, p(\beta)}{p(y)} \qquad (9)$$

where $p(y) = \int p(y \mid \beta) p(\beta) d\beta$ is a normalizing constant which ensures the posterior pdf integrates to unity. Hence,

$$p(\beta \mid C(X), y) \propto p(C(X), y \mid \beta)\, p(\beta) \qquad (10)$$

Thus, to apply the Bayesian inference to logistic regression model, we need to combine the data likelihood and an appropriate prior distribution to compute quantities (mean and variance) that summarizes the posterior distribution. Accordingly, we specified the likelihood function for the Bayesian logistic model. In this paper we considered a student *t*-prior with three (3) degrees of freedom, location parameter zero (0) and scale parameter one (1) on both the intercept and coefficients of the model. The choice of this prior was motivated by the studies of Lange et al. [1989] which opined that flat priors allow for robust inference. Besides, the student *t*-prior distribution with location 0, $v$ degree of freedom and scale $\delta$, is considered to constrain the parameter values to lie in a reasonable range, since minimal prior knowledge is provided [Raftery, 1996]. The general student *t*-prior distribution with $v$ degree of freedom, location parameter $\mu$, and scale parameter $\gamma$ is of the form

$$p_v(z) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\Gamma\left(\frac{v}{2}\right)\sqrt{v\pi\gamma}} \left[ \left( 1 + \frac{1}{v} \left( \frac{z-\mu}{\gamma} \right) \right)^2 \right]^{-\Gamma\left(\frac{v+1}{2}\right)} \qquad (11)$$

Applying (11) to each of the regression coefficients, $\beta_j$, with the generic parameters, we have

$$p_{v_j}(\beta_j) = \frac{\Gamma\left(\frac{v_j+1}{2}\right)}{\Gamma\left(\frac{v_j}{2}\right)\sqrt{v_j\pi\delta_j}} \left[ \left( 1 + \frac{1}{v_j} \left( \frac{\beta_j - \mu_j}{\delta_j} \right) \right)^2 \right]^{-\Gamma\left(\frac{v_j+1}{2}\right)} \qquad (12)$$

For our application with the specified parameter values for the priors on the $\beta$s, the prior model (13) reduces to


DOI: 10.105515/JBE.40.3.41

$$P_{vj}\left(\beta_j\right)=\frac{1}{\sqrt{3\pi}\left(1+\frac{\beta_j^2}{3}\right)^2} \tag{13}$$

Based on the specified likelihood function in (7) and the prior distribution (13), the posterior distribution can be expressed as

$$p\left(\beta_j \mid C\left(X\right), y\right)=\left\{\prod_{i=1}^{m}\left[\frac{e^{C(X_i)\beta}}{1+e^{C(X_i)\beta}}\right]^{y_i}\left[\left(1+e^{C(X_i)\beta}\right)^{-1}\right]^{1-y_i}\right\}\left\{\frac{1}{\sqrt{3\pi}\left(1+\frac{\beta^2}{3}\right)^2}\right\} \tag{14}$$

Unfortunately, the posterior distribution in (14) has no closed form, hence, it is difficult to derive the marginal posterior corresponding to each coefficient by integration due to intractability. Therefore, approximations need to be applied to find an analytical expression for the posterior [see, Drugowitsch, 2019]. Accordingly, various techniques, for instance, stochastic techniques such as MCMC methods and deterministic techniques such as Variational Bayes methods exist in statistical literature for approximating posterior distributions. We adopt the MCMC approach to inference.

**5.2 MCMC Inference for Binary logistic regression**

This section focuses on exact Bayesian inference via the Markov Chain Monte Carlo method (MCMC). It is sometimes difficult to handle resulting posterior distributions from an assumed Bayesian model. This happens if the assumed model exhibits some complex structure based on the data likelihood and the joint prior probability models for all unknown parameters. For highly complex joint Bayesian posteriors, it is often hard to obtain the marginal parameter-wise posterior distributions for the development of an appropriate parameter inference scheme within the MCMC paradigm, since MCMC depends on the full conditionals of the parameters involved in a given model. It is straightforward to see that if the marginal posteriors of parameters in a given model can be obtained in closed or standard forms, it will motivate the use of a very simple MCMC method. In particular, when marginal posteriors of all parameters are available and can be identified in standard distributional form, then the appropriate MCMC method applicable is the Gibbs sampler. However, for parameters with no closed-form marginal posterior distributions, a general approach to inference is via the Metropolis-Hasting sampling algorithm. It may happen in some scientific problems for which the assumed statistical model considered within the Bayesian may have one or more parameters yielding closed-form marginal posteriors while others may not have marginal posteriors that can be identified to have some standard distributional form. In such a situation, there exists a mixture of closed-form and non-closed form marginal posteriors of parameters. An appropriate inference scheme for such problems must consider Gibbs steps for parameters with closed-form marginal posteriors and Metropolis-Hasting steps for parameters with


DOI: 10.105515/JBE.40.3.41

non-closed form marginal posteriors leading to a hybrid MCMC scheme [Hastings, 1970, Metropolis et al., 1953]. In particular, when the Metropolis-Hasting steps are embedded within the Gibbs steps, the resulting MCMC sampler is termed Metropolis-within-Gibbs sampler [Givens and Hoeting, 2013]. The hybrid type of MCMC samplers belong the advanced MCMC methods and readers are referred to Givens and Hoeting [2013], Gelman et al. [2013], and Andrieu and Thoms [2008] for more details on MCMC methods and its advanced counterparts.

Now considering our binary logistic regression model and assessing the nature of the marginal posteriors associated with the parameters, we observe that regression parameters $\beta$'s all yield marginal posterior distributions that have no closed form. This condition presented by the marginal posteriors motivates the use of the Metropolis-Hasting method for parameter inference. As result, we develop a problem-specific Metropolis-Hasting (MH) sampler for inference. The inference algorithm is outlined in Algorithm 1.

**Algorithm 1** Metropolis-Hasting algorithm

1: Initialization: Set MCMC sample size, $\mathcal{N}_m$,

2: Set $t = 0$ and select starting values $\beta_j^{[0]}, j = 1, \ldots, p$.

3: Sample $\beta_j^{[t+1]}$ based on the following steps

1. Sample $\beta_j^{\star}$ from $g\left(\cdot \mid \beta_j^{[t]}\right)$

2. Choose $\beta_j^{[t+1]}$ based on:

$$\beta_j^{[t+1]} = \begin{cases} \beta_j^{\star} & \text{with probability min } \left\{\omega\left(\beta_j^{[t]}), \beta_j^{\star}\right), 0\right\} \\ \beta_j^{[t]} & \text{otherwise,} \end{cases}$$

where

$$\omega\left(\beta_j^{[t]}), \beta_j^{\star}\right) = \left[\frac{p\left(\beta_j^{\star} | \beta_j^{[t]}, c(X_j), y_i\right) g\left(\beta_j^{[t]} \mid \beta_j^{\star}\right)}{p\left(\beta_j^{[t]} \mid \beta_j^{\star}, c(X_j), y_i\right) g\left(\beta_j^{\star} \mid \beta_j^{[t]}\right)}\right],$$

4: Set $t = t + 1$

5: If $t < \mathcal{N}_m$ repeat step 3. Otherwise stop.

## 6 Performance evaluation

In this paper, we considered some performance measures that existed in literature for the evaluation of the proposed methods. The performance of predictive models is based on the


DOI: 10.105515/JBE.40.3.41

accuracy, sensitivity, specificity derived from the confusion matrix as shown in Table 2, as well as the area under (AUC) the receiver-operating-characteristics (ROC) curve.

Table 2: Confusion matrix

| | | Predicted situation | |
|---|---|---|---|
| | | Negative (0) | Positive (1) |
| Actual situation | Negative (0) | $T_rN$ | $F_sP$ |
| | Positive (1) | $F_sN$ | $T_rP$ |

In particular, the accuracy measures the proportion of correctly classified predictions, thus, the ratio of correct classifications captured to the total cases (subjects). We defined accuracy as:

$$Accuracy = \frac{T_rP + T_rN}{T_rP + F_sP + F_sN + T_rN}$$

Sensitivity also known as recall on the other hand is a measure of the proportion of correctly classified positive cases (subjects). It is derived as

$$Sensitivity = \frac{T_rP}{T_rP + F_sN}$$

Specificity also known as precision is a measure of the proportion of correctly classified negative cases (subjects). It is derived as

$$Accuracy = \frac{T_rN}{T_rN + F_sP}$$

Another performance measure considered is the receiver operating characteristics (ROC). It is a unified plot of the proportion of correctly classified positive cases (sensitivity) against the proportion of the incorrectly classified negative cases (1-specificity) for all possible thresholds (ranges between 0 and 1). This plot aids in the evaluation of a classifier's performance based on the area under the curve (AUC) which is between 0 and 1. With regards to AUC, the classification algorithm is deemed good or able to discriminate when the area under the curve is large or above the diagonals [Fern´andez et al., 2018, Giancristofaro and Salmaso, 2003].

**7 Data description**

This section provides a brief overview of the Pima Indian diabetes dataset sourced from the UCI repository of machine learning for this study. This data of size 768 by 9 is multivariate in nature, comprising the binary response and a set of predictors collected on eight medical attributes of suspected diabetic patients from Arizona, USA. The response in this dataset is whether a patient tested positive for diabetes or not, whereas the covariates considered were two patient-specific covariate - number of times being pregnancies (PRG) and age (in years), as well as six diabetes


DOI: 10.105515/JBE.40.3.41

disease diagnostic criteria covariates namely glucose (GLU), diastolic blood pressure (DBP), body mass index (BMI), insulin (INS), skinfold thickness
(SFT), and diabetes pedigree function (DPF). The summaries of the dataset are provided in Table3.

**7.1 Data pre-processing**

The datasets were pre-processed to make it more productive in enhancing the fitting performance of models, due to the presence of missingness [zero (0) values] within the covariate or predictor space as revealed by the summary statistics of the original dataset in Table 4.

Table 3: Summary statistics of covariates related to patient's characteristics and diabetic diagnosis

| Covariates | Statistic | | | | | | |
|---|---|---|---|---|---|---|---|
| | Min | Max | Mean | Median | Mode | SD | Skewness |
| Age | 21.00 | 81.00 | 33.24 | 29.00 | 22.00 | 11.76 | 1.13 |
| PRG | 0.00 | 17.00 | 3.85 | 3.00 | 1.00 | 3.37 | 0.90 |
| GLU | 0.00 | 199.00 | 120.89 | 117.00 | 99.00 | 31.97 | 0.17 |
| DBP | 0.00 | 122.00 | 69.11 | 72.00 | 70.00 | 19.36 | -1.84 |
| SFT | 0.00 | 99.00 | 20.54 | 23.00 | 0.00 | 15.95 | 0.11 |
| INS | 0.00 | 846.00 | 79.80 | 30.50 | 0.00 | 115.24 | 2.27 |
| BMI | 0.00 | 67.10 | 31.99 | 32.00 | 32.00 | 7.88 | -0.43 |
| DPF | 0.08 | 2.42 | 0.47 | 0.37 | 0.25 | 0.33 | 1.92 |

In particular, since medical results cannot be zero (0), all minimum value of zero (0) for diabetes diagnostic covariates were imputed by the mean of these covariates (see Table 2). Besides, in line with the previous studies, the pregnancies variable was transformed into a 0/1 nominal feature, with 1 being previously pregnant and 0 being was never pregnant, since number of pregnancies has no association with diabetes [see, for instance, Patil et al., 2010].

Table 4: Summary statistics of processed continuous covariate data

| Covariates | Statistic | | | | | | |
|---|---|---|---|---|---|---|---|
| | Min | Max | Mean | Median | Mode | SD | Skewness |
| Age | 21.00 | 81.00 | 33.24 | 29.00 | 22.00 | 11.76 | 1.13 |
| GLU | 44.00 | 199.00 | 121.69 | 117.00 | 99.00 | 30.44 | 0.53 |
| DBP | 24.00 | 122.00 | 72.41 | 72.21 | 70.00 | 12.10 | 0.14 |
| SFT | 7.00 | 99.00 | 29.15 | 29.15 | 29.15 | 8.79 | 0.82 |
| INS | 14.00 | 846.00 | 155.55 | 155.55 | 155.55 | 85.02 | 3.02 |
| BMI | 18.20 | 67.10 | 32.46 | 32.40 | 32.00 | 6.88 | 0.60 |
| DPF | 0.08 | 2.42 | 0.47 | 0.37 | 0.25a | 0.33 | 1.92 |


DOI: 10.105515/JBE.40.3.41

## 8 Implementation of methods

The implementation of the proposed methods and algorithms were conducted using R statistical software. In particular, all codes were written in R and run on an Intel (R) Core (TM) i3–7020U CPU @ 2.30GHz workstation. Note that the construction of the robust covariates requires the estimation of the underlying probability density function of the data. As result, its implementation considered the kernel density estimation procedure implemented in the R package *ks* [Duong et al., 2007]. The kernel density estimator of $f\left(x_j\right)$ based on a random sample of original covariates $x_1, x_2, \ldots, x_m$ is defined as

$$\hat{f}\left(x_j\right)=\frac{1}{mh}\sum_{i=1}^{m}K\left(\frac{x_j-x_{jt}}{h}\right) \qquad (15)$$

where $K$ denotes a kernel and bandwidth $h$ is smoothing parameter [Kankanige and Bailey, 2014]. In the implementation $h$ in (15) was set to the smoothed cross-validation estimator implemented in the R package *ks*. Modeling fitting under the classical approach was done via the generalized linear model package, *glm* R. However, the Bayesian modeling fitting based on the Metropolis-Hasting algorithm was implemented using the MCMC package, *MCMC pack* in R [Martin et al., 2011]. In all the experimentation, the appropriate initialization utilized in the associated packages were considered.

## 9 Examples

We evaluate the performance of the proposed methods using simulation and real data application. Two sets of simulations are considered based on the real data pattern. First simulation focuses on generation of synthetic data via perturbation of the original covariates using their out of sample means and variances. In the second simulation, the perturbation is tailored towards varying the variance in the first simulation. This is to check if the observed performance is not sensitive to the variance.

### 9.1 Example 1: Simulation 1

In this example, we consider data generated in line with the real Pima Indian Diabetes dataset. The real data is perturbed appropriately within the predictor space based on the underlying features of the data. The rationale is to introduce some outlying observations across predictors at varying levels in order to access the performance of the proposed methods. Let $\delta$ denotes the level of perturbation of interest. We set the candidates for $\delta$ to be within (1% - 50%) in the predictor space. For each $\delta$ value, corresponding percentage of predictors were randomly replaced with new candidates generated as follows.

$$X_j=\mu_{X_j}+\nu_j,\quad \nu_j\sim \mathrm{N}\left(0,\ \sigma^2_{X_j}\right), \qquad (16)$$

where $\mu_{X_j}$ denotes the out of sample mean and $\sigma^2_{X_j}$ is the variance. Using the above generation procedure, a synthetic diabetes dataset of the same size as the real dataset was generated. The proposed methods were then implemented using the simulated data. The results of the implementation are presented next.


DOI: 10.105515/JBE.40.3.41

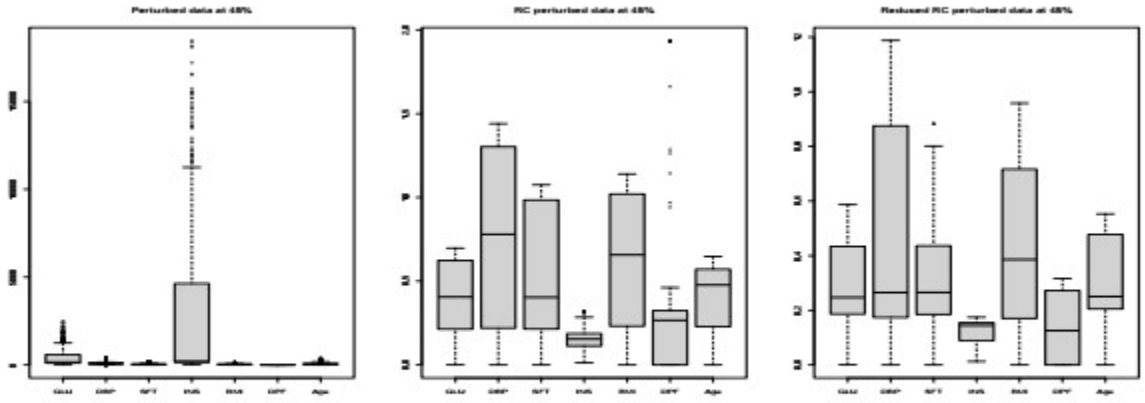

Figure 1: Boxplot of perturbed dataset at 45% under simulation 1.

Figure 9.1 shows the nature of the predictor information in terms of boxplot for the simulated dataset with 45% outlying observations introduced. The first, second and third plots are respectively the simulated predictor data, full robust covariate data, and reduced robust covariate data. Clearly, some significant differences among the boxplots are evident with most outlying observations automatically controlled.

Table 5: Simulation 1: Predictive performance statistics

| Accuracy | | | | | | |
|---|---|---|---|---|---|---|
| | Classical Approach | | | MCMC Approach | | |
| | Perturbed | RC | RC | Perturbed | RC | RC |
| $\delta$ | data | Contribution | Reduced | data | Contribution | Reduced |
| 1% | 0.75 | 0.67 | 0.98 | 0.76 | 0.67 | 1.00 |
| 5% | 0.73 | 0.68 | 0.98 | 0.70 | 0.67 | 0.98 |
| 10% | 0.72 | 0.68 | 0.98 | 0.69 | 0.67 | 0.99 |
| 15% | 0.72 | 0.67 | 1.00 | 0.69 | 0.67 | 1.00 |
| 20% | 0.68 | 0.67 | 0.99 | 0.67 | 0.65 | 0.99 |
| 25% | 0.71 | 0.66 | 0.99 | 0.67 | 0.66 | 0.98 |
| 30% | 0.70 | 0.66 | 1.00 | 0.68 | 0.66 | 0.97 |
| 35% | 0.71 | 0.67 | 0.99 | 0.67 | 0.66 | 0.99 |
| 40% | 0.69 | 0.67 | 1.00 | 0.66 | 0.66 | 0.99 |
| 45% | 0.67 | 0.66 | 0.99 | 0.66 | 0.66 | 0.99 |
| 50% | 0.66 | 0.64 | 1.00 | 0.66 | 0.66 | 0.99 |



DOI: 10.105515/JBE.40.3.41

Table 5 reports the predictive accuracy results of the proposed methods under both classical and Bayesian inference methods for first simulated example. It can observe that the presence of outliers can impair the predictive performance of the binary logistic regression for diabetes mellitus prediction regardless of the modelling framework. In particular, the results based on the perturbed data at all levels of $\delta$ (1% - 50%) clearly shows that some level of fluctuation exist in the predictive accuracy. However, the result based on the robust covariate (RC) contribution indicates that some level of consistency in predictive accuracy was ensured due to the automatic control of outliers. This therefore suggest that the high performance seen in the perturbed data was actually due to outliers, hence, the reduce performance under the RC contribution was as a result of the control. Furthermore, the result based on the RC reduced data shows that the reduction has offered significant improvement in predictive performance over the overall data with robust covariate contribution. The implication is that, though outliers have been controlled, not all data points are useful for modelling and predicting diabetes mellitus.

**9.2 Example 2: Simulation 2**

This simulation follows the settings in simulation 9.1 with some modification. The modification is in the direction of the variance. Here, predictors are replaced with assumed level, $\delta$ using

$$X_j = \mu_{X_j} + \nu_j,\ \nu_j \sim N\left(0,\ \sigma^2_{X_j}\right), \qquad (17)$$

where $\mu_{Xj}$ is as in example 9.1 and $\sigma^2_{X_j} = \sqrt{X_j}$. Also, a synthetic diabetes dataset of the same size as the real dataset was simulated and used to implement the proposed methods. The results of implementation of simulation study 2 are presented next.

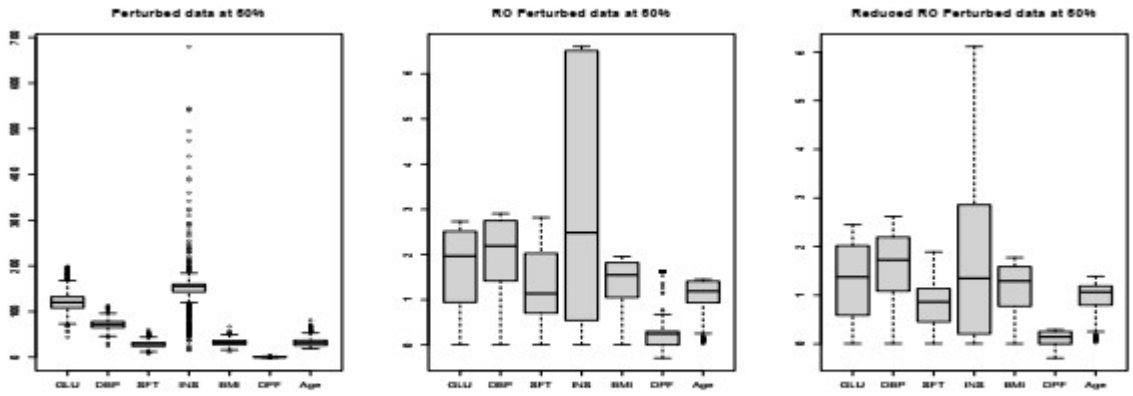


Figure 2: Boxplot of perturbed dataset at 50% under simulation 2.

Figure 9.2 shows the distribution of covariate information for dataset generated in simulation 2, with about 50% outlying observation introduced. Similarly, the plot exhibits same pattern as established in Figure 9.1. Thus, some significant differences among the boxplots are evident with most outlying observations automatically controlled.


DOI: 10.105515/JBE.40.3.41

Table 6: Performance of proposed method under simulation 2

| Performance Accuracy | | | | | | |
|---|---|---|---|---|---|---|
| | Classical Approach | | | MCMC Approach | | |
| | Perturbed | RC | RC | Perturbed | RC | RC |
| $\delta$ | data | Contribution | Reduced | data | Contribution | Reduced |
| 1% | 0.77 | 0.66 | 1.00 | 0.77 | 0.66 | 1.00 |
| 5% | 0.78 | 0.66 | 0.99 | 0.77 | 0.66 | 0.99 |
| 10% | 0.75 | 0.66 | 0.98 | 0.75 | 0.65 | 0.99 |
| 15% | 0.74 | 0.65 | 1.00 | 0.73 | 0.65 | 0.99 |
| 20% | 0.75 | 0.66 | 0.98 | 0.75 | 0.65 | 0.99 |
| 25% | 0.70 | 0.65 | 0.98 | 0.72 | 0.66 | 0.97 |
| 30% | 0.71 | 0.64 | 0.98 | 0.72 | 0.66 | 0.99 |
| 35% | 0.69 | 0.65 | 0.99 | 0.71 | 0.65 | 1.00 |
| 40% | 0.70 | 0.64 | 0.99 | 0.70 | 0.66 | 0.99 |
| 45% | 0.69 | 0.66 | 0.99 | 0.70 | 0.65 | 0.98 |
| 50% | 0.65 | 0.64 | 0.96 | 0.70 | 0.65 | 0.99 |

Table 6 presents results that aids the evaluation of the proposed method under simulation 2 via two modeling algorithms – classical and Bayesian approaches. The results thus shows that although the application of both the classical and Bayesian logistic algorithm to the robust covariates contribution data did not show any improvement in the predictive accuracy of the model over the perturbed data application at all levels of perturbation ($\delta$), it could be observed that the application of these two classification algorithms to the RC reduced dataset resulted in enhanced predictive accuracy of the model at all levels.

### 9.3 Example 3: Application to Pima Indian Diabetes data

This section focuses on the results and discussion of application of the developed methods on the real Pima Indian Diabetes dataset. In particular, attention centered on the predictive performance in terms of the measures outlined in Section 6.


DOI: 10.105515/JBE.40.3.41

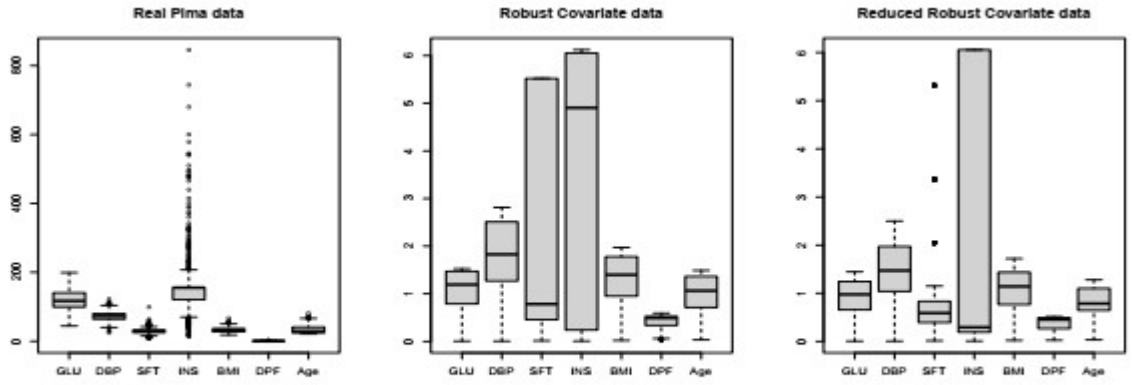


Figure 3: Boxplot of methods application with the real dataset.

Figure 9.3 displayed the distribution of covariates information based on the pre-processed Pima diabetes dataset and data generated using both the full and reduced robust covariate. Clearly, the dynamics of the plot shows that although some level of outlying observations exists across multiple covariates in the plot of the real Pima dataset, there is evidence of these outliers being automatically controlled as seen in the plots of both the full robust covariate and reduced robust covariate data. Subsequently, the predictive performance of the proposed method is examined in both the Classical and Bayesian inference paradigms. Table 7 presents the predictive performance of the methods using the real Pima Indian Diabetes dataset.

Table 7: Performance of Method Under MCMC for the three datasets

| | Performance Metrics | | | | | |
|---|---|---|---|---|---|---|
| | Classical Approach | | | Bayesian Approach | | |
| Dataset | AC | SN | SP | AC | SN | SP |
| Real data | 0.77 | 0.58 | 0.87 | 0.78 | 0.92 | 0.44 |
| RC Contribution | 0.66 | 0.30 | 0.85 | 0.67 | 0.93 | 0.14 |
| RC Reduced | 0.99 | 1.00 | 0.99 | 0.99 | 0.99 | 1.00 |

*AC=Accuracy; SN=Sensitivity; SP=Specificity*

Table 7 report the predictive accuracy for proposed method's application with real Pima Indian diabetes dataset under both classical and Bayesian logistic classifier. Generally, results under the Bayesian (MCMC) approach seems to be similar to that obtained under classical approaches, however, there is a marginal difference in the predictive accuracy under these two computational approaches. In particular, the result showed an improved accuracy of 78% and 67% under the MCMC approach for both real Pima data and robust covariate (RC) dataset respectively as compared to the classical approach which attained about 77% and 66% accuracy respectively. The implication is that, under the classical approach, the parameters are fixed but unknown, however,


DOI: 10.105515/JBE.40.3.41

under the MCMC approach, the parameters are assumed random and the uncertainties around how these parameters are quantified and calibrated via the prior information enabled an improvement in the predictive accuracy under the MCMC approach. Additionally, the results show an equal predictive performance, thus, 99% accuracy, under both classical and MCMC approaches for the RC reduced data. This finding suggest that predictors (covariates) are very critical in logistic regression, such that if an outlier exist within the predictor space, it can impair the performance regardless of the modelling framework, however, predictive performance remain same when controlled using density robust covariate and there is some sort of data reduction (thus, when unnecessary information is discarded).

We next present the performance of the proposed method using the receiver-operating-characteristics curve under both classical and Bayesian approaches and the result is presented in Figure 6 and Figure 7 respectively.

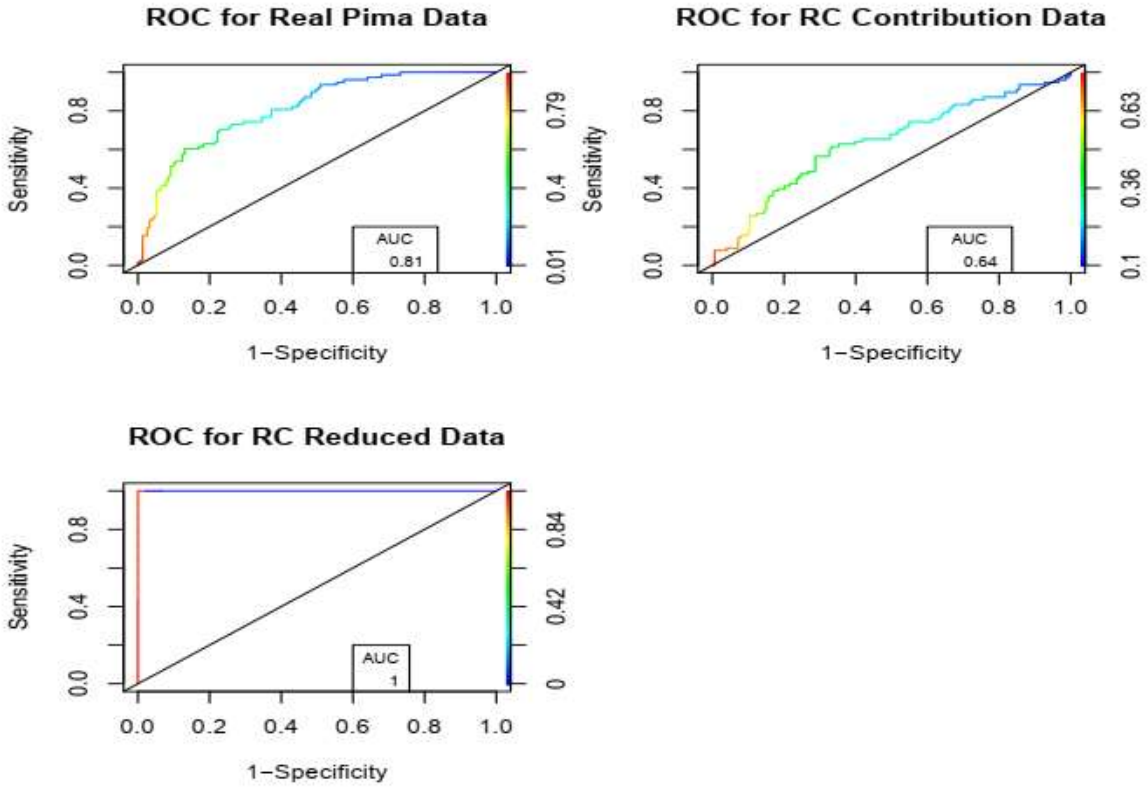


Figure 4: ROC curve for Real, Contribution and Reduced Dataset for LR under MLE.


DOI: 10.105515/JBE.40.3.41

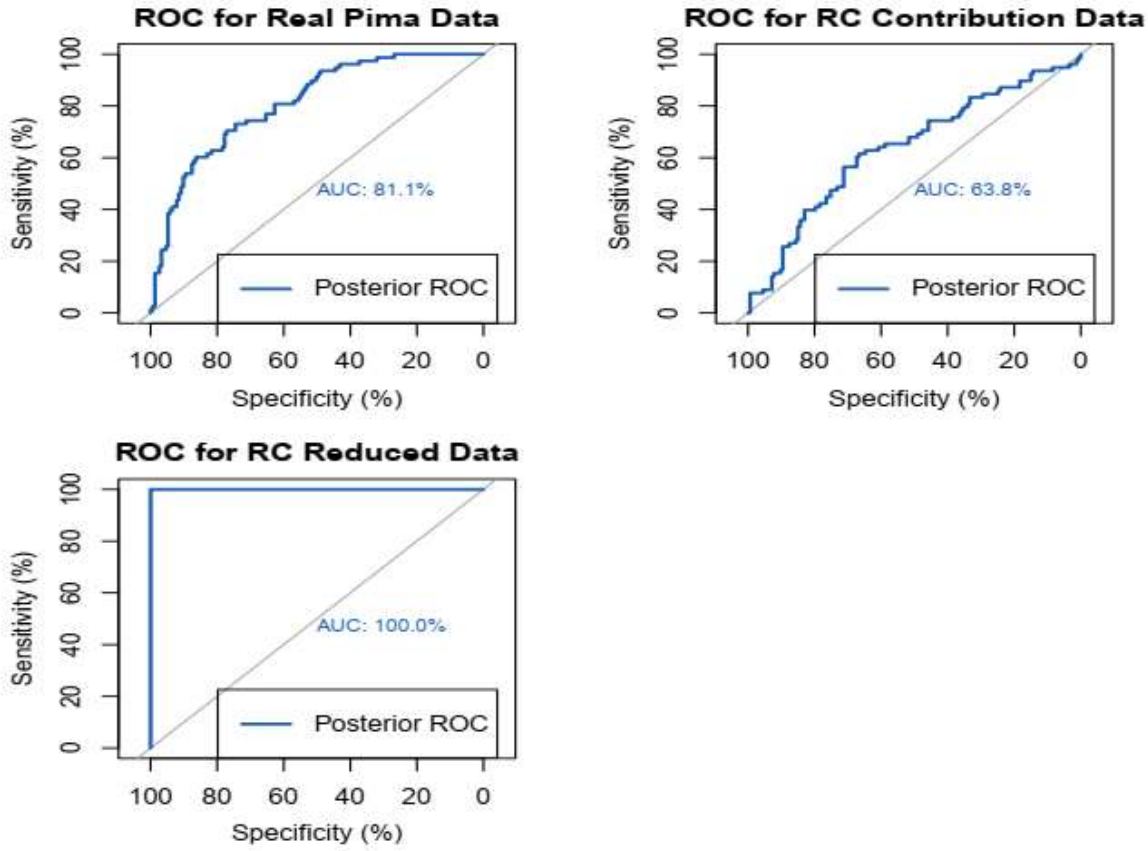


Figure 5: ROC curve for Real, Contribution and Reduced Dataset for Bayesian LR.

Comparatively, the result in Figure 6 and Figure 7 suggests that the area under the ROC curve under both classical and Bayesian application with the reduced dataset is significantly higher than those attained with the application of the method with the real Pima data and the RC contribution dataset, suggesting an excellent classification accuracy.

**10. Conclusion**

In this paper, we have proposed and implemented a novel approach for improving the predictive performance of Binary logistic regression for Gestational diabetes mellitus prediction using alternative predictors derived from the original predictor data. The alternative predictors are constructed as features from the underlying probability density functions of the original predictor variables via kernel density estimation, using the idea of moments about the origin. With the probability density functions as weights in the alternative predictors, observation-specific contributions underlying the predictors can be automatically assessed, allowing extreme values to be handled appropriately within the probability density space. This allows existing extreme values to be utilized for model building without deleting them as deletion leads to loss of information. In addition, the predictor-specific autocorrelations are treated automatically with the help of the kernel density estimation procedure. In this regard, the alternative predictors become robust to


DOI: 10.105515/JBE.40.3.41

extremes and predictor-specific autocorrelation. The above appealing features allow the development of binary logistic regression models in both the classical and Bayesian framework for public health problems as such modeling and prediction of gestational diabetes mellitus. Appropriate classical and Bayesian binary logistic regression models were developed using the alternative predictors (robust covariates). The usefulness of the developed methods in enhancing the predictive performances of binary logistic models for gestational diabetes mellitus is illustrated using both simulations and real data of the Pima Indian Diabetes data. Most importantly, it was realized from the experimentation using both simulations with varying degrees of outlying observations and real-data application that, the robust covariates allowed for automatic control of extreme values in the predictor spaces and ensured dimension reduction. These features enhanced the performance of the logistic classifier in both the classical and Bayesian computational approaches. In particular, the developed methods established that logistic classification algorithms performed at an improved level when implemented with reduced robust covariates as compared with their performance with the original covariates and the full contributions informed by the original covariates. In comparison with an existing state-of-the-art method that uses the PCA and K-Means clustering [Zhu et al., 2019], based on the Pima Indians diabetes dataset, the proposed methods outperformed it by an order of magnitude especially, with the reduced robust covariates.

**Funding**

Etornam Kwame Kunu's research was fully supported by the Samuel and Emelia BrewButlerSGS/GRASAG, University of Cape Coast Research Grant. The research of the remaining authors was supported by the research component of the Book and Research grant by the Government of Ghana for the 2020-2021 academic year.

**Acknowledgements**

The authors register their profound gratitude to the Government of Ghana for supporting the work of David Kwamena Mensah, Francis Eyiah-Bediako and Samuel Assabil and Samuel and Emelia Brew-Butler for supporting the research of Etornam Kwame Kunu.

**References**

WHO. The top 10 causes of death. https://www.who.int/news-room/fact-sheets/detail/thetop-10-causes-of-death, 2020. Accessed: 2021-02-17. 3

IDF. Gestational diabetes. https://www.idf.org/aboutdiabetes/what-is-diabetes/factsfigures.html, 2020. Accessed: 2021-02-17. 3

Getasew Mulat Bantie, Achenef Almaw Wondaye, Efrem Beru Arike, Mesfin Tenagne Melaku, Simegnew Tilaneh Ejigu, Abel Lule, Wondemagegn Mulu Lingerew, and Koku Sisay Tamirat. Prevalence of undiagnosed diabetes mellitus and associated factors among adult residents of Bahir Dar city, northwest Ethiopia: a community-based cross-sectional study. *BMJ open*, 9(10):e030158, 2019. 3

Ioannis Kavakiotis, Olga Tsave, Athanasios Salifoglou, Nicos Maglaveras, Ioannis Vlahavas,


DOI: 10.105515/JBE.40.3.41

and Ioanna Chouvarda. Machine learning and data mining methods in diabetes research. *Computational and structural biotechnology journal*, 15:104–116, 2017. 3

Ravinder Ahuja, Subhash C Sharma, and Maaruf Ali. A diabetic disease prediction model based on classification algorithms. *Annals of Emerging Technologies in Computing (AETiC), Print ISSN*, pages 2516–0281, 2019. 3

Aishwarya Mujumdar and V Vaidehi. Diabetes prediction using machine learning algorithms. *Procedia Computer Science*, 165:292–299, 2019. 3

A Gelman. Hill j. 2007 data analysis using regression and multilevel/hierarchical models. *Analytical methods for social research. Cambridge University Press*, 2007. 3, 6

Aiswarya Iyer, S. Jeyalatha, and Ronak Sumbaly. Diagnosis of diabetes using classification mining techniques. *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, 2015. 3

Tarun Jhaldiyal and Pawan Kumar Mishra. Analysis and prediction of diabetes mellitus using PCA, REP and SVM. *International Journal of Engineering and Technical Research (IJETR)*, 2(8):164–166, 2014. 3

Han Wu, Shengqi Yang, Zhangqin Huang, Jian He, and Xiaoyi Wang. Type 2 diabetes mellitus prediction model based on data mining. *Informatics in Medicine Unlocked*, 10: 100–107, 2018. 4

Changsheng Zhu, Christian Uwa Idemudia, and Wenfang Feng. Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. *Informatics in Medicine Unlocked*, 17:100179, 2019. 4, 28

Kenneth Chukwuemeka Asanya, Mohamed Kharrat, Akaninyene Udo Udom, and Emmanuel Torsen. Robust bayesian approach to logistic regression modeling in small sample size

utilizing a weakly informative student'st prior distribution. *Communications in StatisticsTheory and Methods*, pages 1–11, 2021. 6

Alan Agresti. *Foundations of Linear and Generalized Linear Models*. John Wiley & Sons, 2015. 10

Kenneth L. Lange, Roderick J. A. Little, and Jeremy M. G. Taylor. Robust statistical modeling using the t distribution. *Journal of the American Statistical Association*, 84 (408):881–896, 1989. 11

Adrian E. Raftery. Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika*, 83(2):251–266, 1996. 11

Jan Drugowitsch. Vblinlogit: Variational bayesian linear and logistic regression. *Journal of Open Source Software*, 4(38):1359, 2019. 12

W. K. Hastings. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, pages 97–109, 1970. 13

Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953. 13


**DOI**: 10.105515/JBE.40.3.41

G.H. Givens and J.A. Hoeting. *Computational Statistics*. Wiley Series in Computational Statistics. Wiley, 2013. ISBN 9780470533314. URL https://books.google.com. gh/books?id=bCJx53VQS7IC. 13

Andrew Gelman, John B. Carlin, Hal S Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2013. 13

Christophe Andrieu and Johannes Thoms. A tutorial on adaptive mcmc. *Statistics and computing*, 18(4):343–373, 2008. 14

Alberto Fern´andez, Salvador Garc´ıa, Mikel Galar, Ronaldo C. Prati, Bartosz Krawczyk, and Francisco Herrera. *Learning from imbalanced data sets*, volume 10. Springer, 2018. 16

Rosa Arboretti Giancristofaro and Luigi Salmaso. Model performance analysis and model validation in logistic regression. *Statistica*, 63(2):375–396, 2003. 16

Bankat M Patil, Ramesh Chandra Joshi, and Durga Toshniwal. Hybrid prediction model for type-2 diabetic patients. *Expert systems with applications*, 37(12):8102–8108, 2010. 17

Tarn Duong et al. ks: Kernel density estimation and kernel discriminant analysis for multivariate data in r. *Journal of Statistical Software*, 21(7):1–16, 2007. 18

Yamuna Kankanige and James Bailey. Improved feature transformations for classification using density estimation. In *Pacific Rim International Conference on Artificial Intelligence*, pages 117–129. Springer, 2014. 18

Andrew D. Martin, Kevin M. Quinn, and Jong Hee Park. MCMCpack: Markov Chain Monte Carlo in R. *Journal of Statistical Software*, 42(i09), 2011. 18


**DOI**: 10.105515/JBE.40.3.41