Eur. J. Stat. 2 (2022) 8
doi: 10.28924/ada/stat.2.8

Multinomial Naïve Bayes Classifier: Bayesian versus Nonparametric Classifier Approach

Rasaki Olawale Olanrewaju[1], Sodiq Adejare Olanrewaju[2,*], Lukman Abiodun Nafiu[3]

[1]*Department of Mathematical Sciences, Pan African University Institute for Basic Sciences, Technology and Innovation, Nairobi, P.O. Box 62000-00200, Kenya*
olanrewaju_rasaq@yahoo.com
[2]*Department of Statistics, University of Ibadan, Ibadan, 900001, Nigeria*
sodiqadejare19@gmail.com
[3]*Department of Mathematics and Statistics, Kabale University, Kabale, Uganda*
lanafiu@kab.ac.ug
[*]*Correspondence: olanrewaju_rasaq@yahoo.com*

Abstract. This paper proposes a Naïve Bayes Classifier for Bayesian and nonparametric methods of analyzing multinomial regression. The Naïve Bayes classifier adopted Bayes' rule for solving the posterior of the multinomial regression via its link function known as Logit link. The nonparametric adopted Gaussian, bi-weight kernels, Silverman's rule of thumb bandwidth selector, and adjusted bandwidth as kernel density estimation. Three categorical responses of information on 78 people using one of three diets (Diet A, B, and C) that consist of scaled variables: age (in years), height (in cm), weight (in kg) before the diet (that is, pre-weight), weight (in kg) gained after 6 weeks of diet were subjected to the classifier multinomial regression of Naïve Bayes and nonparametric. The Gaussian and bi-weight kernel density estimation produced the minimum bandwidths across the three categorical responses for the four influencers. The Naïve Bayes classifier and nonparametric kernel density estimation for the multinomial regression produced the same prior probabilities of 0.3077, 0.3462, and 0.3462; and A prior probabilities of 0.3077, 0.3462, and 0.3462 for Diet A, Diet B, and Diet C at different smoothing bandwidths.

1. Introduction

Application of probabilistic based classification to Bayes' theorem with vigorous independency assumptions is referred to as Naïve Bayes classifier. It is a direct descriptive approach for re-modifying probability model to be an unconstrained attribute model (Wibawa *et al.*, 2019). Elaborately, a Naïve Bayes classifier formally presupposes that the absence (or presence) of a particular attribute of a category is unrelated to the absence (or presence) of other traits. For instance, a fruit might be called a banana if it is yellow (or green for unripe one), has a

Received: 9 Jan 2022.
Key words and phrases. kernel density estimation; bandwidth; Bayes' rule; multinomial regression; Naïve Bayes classifier.

curved shape, and contains good amount of calories, 15-17 cm in length and 42" thereabout in diameter. The mentioned attributes depend on each other, upon the subsistence, varieties, or other attributes; what Naïve Bayes classifier does is to consider all these attributes to be independent as contributors to the likelihood that the fruit is a banana (Zhang & Gao, 2011; Salmi & Rustam, 2019).

Depending on the accuracy of the probability model, Naïve Bayes classifier can be either efficiently trained or untrained in a supervised learning setting. In many pragmatic applications, parameter estimation of Naïve Bayes models is usually via maximum likelihood. However, one can work with Naïve Bayes model without knowing the prior probabilistic distribution or using any Bayesian methods (Poovaraghan *et al.*, 2019).
It is of utmost importance in machine learning algorithms and regression analysis to estimate either the unknown probabilistic function (either probability density function or probability mass function) or attached probabilistic function to a given dataset based on sound and convincing subjectivity, e.g. Bayesian classifiers, mutual information-based attribute selection algorithms, and density-based clustering algorithms. In instances where probabilistic function is unknown, Kernel Density Estimator (KDE) is thoroughly constructed in advance (Xu, 2018).

KDE is a nonparametric (distribution-free) technique of statistical modeling that makes use of data or observations to build a statistical model. In other words, it is a technique for estimating probabilistic function for a given data without any circumstance probabilistic function. The probabilistic function (be it probability density function or probability mass function) of a given dataset can be obtained via commingling kernel functions, usually via begetting each value in the given dataset. However, KDE is an effective way of estimating probabilistic function of a given data when the distribution of the data is unknown. In other words, KDE makes it easy to solve the problem of non-Gaussian distribution when the dataset is of continuous or numeric attributes (Kaviani & Sunita, 2017; Chen *et al.*, 2021).

KDE training method superposes kernel functions in its manifolds with flexible bandwidth to fit the unknown probabilistic function. The most commonly used kernels are Gaussian kernel, triangular, Epanechnikov, biweight, and triweight. In comparison to kernel, bandwidth plays a significant part in the estimation of the distribution of the probabilistic functions: a small bandwidth usually leads to under-smoothed estimation while a large bandwidth usually leads to over-smoothed estimation (Kelly & Johnson, 2021). The Naïve Bayes classifier and the non-parametric KDE will be a framework of multinomial regression analysis in this work. It will be a framework for responses of more than two categorical acumen (multinomial distributed responses) with associated covariates of any form of data measurement. In other words, this research will be engineered by multinomial

regression Naïve Bayes classifier for Bayesian and nonparametric KDE for its responses of more than two categories.

2. Literature Review

Multinomial regression model has been widely used in regression modeling, inter alia, epidemiological and biostatistics bailiwicks. Sarrias & Daziano (2017) anticipated that embedded coefficients in multinomial regression analysis are usually estimated via Maximum Likelihood (ML) or Ordinary Least Square (OLS) on the bases of random sampling or randomization. They advocated that Bayesian multinomial regression analyzes would be more appropriate when samples are generated for outcomes or when the data do not follow randomization (Blizzard *et al.*, 2007). The classical multinomial regression provides the standard penalized ML solutions to multi-class categorical outcomes of a dataset, in contrast to logistic regression that provides solution for dichotomous-type responses (Croissant, 2020).

Nandram (2021) proposed a three-stage hierarchical Bayesian multinomial-Dirichlet-Dirichlet-model for multinomial counts in order to recital for heterogeneity effect. The main contribution of the proposed model was to develop a joint posterior density for the multinomial-Dirichlet–Dirichlet model via a combination of non-parametric and parametric, on bases of nested error regression with the use of iterative re-weighted least squares. According to Johndrow *et al.* (2019), Bayesian estimation for categorical responses with augmented and imbalanced data usually resulted in inefficient sampling behavior. Andrea & Nicola (2014) and Wioletta (2015) came-up with an informative prior of lognormal, Log-F or Gaussian for estimating posterior parameters in multinomial regression analyzes. They affirmed that there has been diverse range of prior distributions (informative or non-informative) with no clear-cut has to the best prior that best sourced a data and parameter(s) to be estimated in a Bayesian setting. According to Mark & David (2000) and Nandram *et al.* (2018), informative priors are pre-knowledge based about the problem based data; such that elicit opinion of experts are usually used to construct antecedent distribution that appropriately premeditates beliefs about the unknown parameter(s). They are of the belief that informative priors (via its subdivision of conjugate and non-conjugate priors) seem to be over-subjective and unscientific. Examples of distributions that provide informative priors are Jeffrey's prior, Inverse-gamma, Walshart, uniform etc. On the other hand, Chen & Fu (2018) affirmed that non-informative priors or diffuse priors are positively biased when bounded by range of distributions less than four. It is based on this that some prominent Bayesian analysts concluded that non-informative priors are misleading, vague and diffuse priors. Examples of non-informative priors are binomial, Poisson, Laplace, and Beta distributions.

In other to bridge the lacuna between informative and non-informative prior, a Naïve Bayes classifier theoretical framework via Baye's rule and nonparametric Kernel Density Estimation (KDE) will be used to analyze multinomial regression. The multinomial regression enables categorical responses (responses of more than two categories) with different types of data measurement for the associated covariates/independent variables. The adopted Baye's rule will be in terms of posterior, prior, normalization and required data as proposed by the English Statistician, Thomas Bayes in 1973, while the nonparametric method will use Gaussian and bi-weight has KDE.

3. A Paradox of Multinomial Regression

According to Sinharay (2010), multinomial distribution is a multivariate colligation of the binomial distribution. Assuming we have independent n-trials with possibility of k^{th} finite outcomes, then the associated probabilities are $\delta_1, \delta_2, \cdots, \delta_k$, such that, $\delta_i \geq 0, \ \ i = 1, \cdots, k$, $\sum_{i=1}^{k} \delta_i = 1$. Then $X = (X_1, X_2, \cdots, X_k)$ follows a multinomial distribution with parameters "δ" and "n", such that, $\delta = (\delta_1, \delta_2, \cdots, \delta_k)$. The Probability Mass Function (PMF) of the multinomial distribution is:

$$P(X_1 = x_1, \cdots, X_k = x_k) = \frac{n!}{x_1! x_2! \cdots x_k!} \delta_1^{x_1} \delta_2^{x_2} \cdots \delta_k^{x_k} \tag{1}$$

It is to be noted that $\sum_{i=1}^{k} x_i = n$. If $k = 2$, the multinomial distribution will be tantamount to binomial distribution. Logit link is the link function of multinomial distribution. It is peculiar to probabilities associated to observations with possible outcomes (e.g. the outcome of either 1, 2, 3, 4, 5 or 6 shows-up when a die is tossed). In situation where possible outcomes are greater than two, multiple probabilities are assigned to observations such that individual probability takes values between zero and one, the sum of the probabilities must be less than or equal to one (Hasan *et al.*, 2016). Assuming there are M-possible outcomes, then there are M-1 probabilities to be estimated via $\left(\delta_i^{[1]}, \cdots, \delta_i^{[M-1]}\right)$. The chance of obtaining the outcome is via $\left(\delta_i^{[M]} = 1 - \sum_{m=1}^{M-1} \delta_i^{[m]}\right)$. The multinomial logit link function (otherwise known as the inverted link function) can be written as:

$$m \log it(\delta_i^{[m]}) = \ln\left(\frac{\delta_i^{[m]}}{\delta_i^{[M]}}\right) = \beta_0^{[m]} + \beta_1^{[m]} x_{i1} + \beta_2^{[m]} x_{i2} + \cdots + \beta_u^{[m]} x_{iu} = X^T \beta_\mu \tag{2}$$

For $i = 1, \cdots, n$, $X^T = (x_{i1}, x_{i2}, \cdots, x_{iu})$, $\beta^T = \left(\beta_1^{[m]}, \beta_2^{[m]}, \cdots, \beta_u^{[m]}\right)$ in matrix form. Where X is the design matrix of $\mu \times \mu$, ρ is the 1 by μ regression coefficients. It is to be noted that the regression coefficients can have different values, that is, the effect of the covariates can be different from their probabilities. The probabilities can be estimated as:

$$\delta_i^{[m]} = \frac{\beta_0^{[m]} + \beta_1^{[m]} x_{i1} + \beta_2^{[m]} x_{i2} + \cdots + \beta_\mu^{[m]} x_{i\mu}}{1 + \sum_{\kappa=1}^{M-1} \exp\left(\beta_0^{[\kappa]} + \beta_1^{[\kappa]} x_{i1} + \beta_2^{[\kappa]} x_{i2} + \cdots + \beta_\mu^{[\kappa]} x_{i\mu}\right)} \tag{3}$$

4. The Multinomial Classification via Gaussian Naïve Bayes (GNB)

The posterior distribution from the GNB assumptions is derived using Baye's rule as follows:

$$P(Y_i/X) = \frac{P(Y_i)P(X/Y_i)}{\sum_{i=1}^{n} P(Y_i)P(X/Y_i)} \quad Y_i = 0, 1, \cdots, n \tag{4}$$

$P(Y_i/X) = Posterior$; $P(X/Y_i) = Data$; $P(Y_i) = Prior$; $\sum_{i=1}^{n} P(Y_i)P(X/Y_i) = Normalization$

Say for

$$P(Y = n/X) = \frac{P(Y = n)P(X/Y = n)}{P(Y = 0)P(X/Y = 0) + P(Y = 1)P(X/Y = 1) + \cdots + P(Y = n)P(X/Y = n)} \tag{5}$$

Dividing both the numerator and denominator by $P(Y = n)P(X/Y = n)$

$$P(Y = n/X) = \frac{\frac{P(Y=n)P(X/Y=n)}{P(Y=n)P(X/Y=n)}}{\frac{P(Y=0)P(X/Y=0)}{P(Y=n)P(X/Y=n)} + \frac{P(Y=1)P(X/Y=1)}{P(Y=n)P(X/Y=n)} + \cdots + \frac{P(Y=n)P(X/Y=n)}{P(Y=n)P(X/Y=n)}} \tag{6}$$

$$P(Y = n/X) = \frac{1}{\frac{P(Y=0)P(X/Y=0)}{P(Y=n)P(X/Y=n)} + \frac{P(Y=1)P(X/Y=1)}{P(Y=n)P(X/Y=n)} + \cdots + \frac{P(Y=n)P(X/Y=n)}{P(Y=n)P(X/Y=n)}} \tag{7}$$

$$P(Y = n/X) = \frac{1}{\frac{P(Y=0)P(X/Y=0)}{P(Y=n)P(X/Y=n)} + \frac{P(Y=1)P(X/Y=1)}{P(Y=n)P(X/Y=n)} + \cdots + 1} \tag{8}$$

$$P(Y = n/X) = \frac{1}{e^{\ln\left[\frac{P(Y=0)P(X/Y=0)}{P(Y=n)P(X/Y=n)}\right]} + e^{\ln\left[\frac{P(Y=1)P(X/Y=1)}{P(Y=n)P(X/Y=n)}\right]} + \cdots + e^{\ln(1)}} \tag{9}$$

$$P(Y = n/X) = \frac{1}{e^{\ln\left[\frac{P(Y=0)P(X/Y=0)}{P(Y=n)P(X/Y=n)}\right]} + e^{\ln\left[\frac{P(Y=1)P(X/Y=1)}{P(Y=n)P(X/Y=n)}\right]} + \cdots + e^{(0)}} \tag{10}$$

$$P(Y = n/X) = \frac{1}{e^{\ln\left[\frac{P(Y=0)}{P(Y=n)} \times \frac{P(X/Y=0)}{P(X/Y=n)}\right]} + e^{\ln\left[\frac{P(Y=1)}{P(Y=n)} \times \frac{P(X/Y=1)}{P(X/Y=n)}\right] + \cdots + 1}} \tag{11}$$

$$P(Y = n/X) = \frac{1}{e^{\left[\ln \frac{P(Y=0)}{P(Y=n)} \times \sum_i \frac{\ln P(X/Y=0)}{\ln P(X/Y=n)}\right]} + e^{\left[\ln \frac{P(Y=1)}{P(Y=n)} \times \sum_i \frac{\ln P(X/Y=1)}{\ln P(X/Y=n)}\right] + \cdots + 1}} \tag{12}$$

$$P(Y = n/X) = \frac{1}{e^{\left[\ln \frac{\delta_0^{[m]}}{\delta_0^{[M]}} \times \sum_i \frac{\ln P(X/Y=0)}{\ln P(X/Y=n)}\right]} + e^{\left[\ln \frac{\delta_1^{[m]}}{\delta_1^{[M]}} \times \sum_i \frac{\ln P(X/Y=1)}{\ln P(X/Y=n)}\right] + \cdots + 1}} \tag{13}$$

The final expression of $P(Y = n/X)$ is in terms of the inverted link function for multinomial distribution. Considering the summations in the denominator of equation (13) and given our assumption of GNB that $P(X_i/Y = y_k)$ is Gaussian, expanding gives:

$$\sum_i \frac{\ln P(X/Y=0)}{\ln P(X/Y=n)} = \sum_i \ln \left[\frac{\frac{1}{\sqrt{2\pi\sigma^2}} e \left(\frac{\left(X_i - \delta_0^{[m]}\right)}{2\sigma_i^2} \right)^2}{\frac{1}{\sqrt{2\pi\sigma^2}} e \left(\frac{\left(X_i - \delta_n^{[m]}\right)}{2\sigma_i^2} \right)^2} \right] \tag{14}$$

$$= \left(\frac{\left(X_i - \delta_n^{[m]}\right)^2 - \left(X_i - \delta_0^{[m]}\right)^2}{2\sigma_i} \right) \tag{15}$$

$$= \sum_i \left(\frac{2X_i \left(\delta_0^{[m]} - \delta_n^{[m]}\right) + \left(\delta_n^{[m]}\right)^2 - \left(\delta_0^{[m]}\right)^2}{2\sigma_i^2} \right) \tag{16}$$

$$= \sum_i \left(\frac{X_i \left(\delta_0^{[m]} - \delta_n^{[m]}\right)}{\sigma_i^2} + \frac{\left(\delta_n^{[m]}\right)^2}{2\sigma_i^2} \right) \tag{17}$$

Equation (17) is called linear weighted sum of the $X_i^T s$. Substituting equation (17) back into equation (13) gives,

$$P(Y=n/X) = \frac{1}{e^{\left[\ln \frac{\delta_0^{[m]}}{\delta_0^{[M]}} \times \sum_i \left(\frac{X_i \left(\delta_0^{[m]} - \delta_n^{[m]}\right)}{\sigma_i^2} + \frac{\left(\delta_n^{[m]}\right)^2}{2\sigma_i^2} \right) \right]} + e^{\left[\ln \frac{\delta_1^{[m]}}{\delta_1^{[M]}} \times \sum_i \left(\frac{X_i \left(\delta_1^{[m]} - \delta_n^{[m]}\right)}{\sigma_i^2} + \frac{\left(\delta_n^{[m]}\right)^2}{2\sigma_i^2} \right) \right] + \cdots + 1}} \tag{18}$$

$$\omega_i = \frac{\left(\delta_0^{[m]} - \delta_n^{[m]}\right)}{\sigma_i^2} + \frac{\left(\delta_n^{[m]}\right)^2}{2\sigma_i^2}; \quad i = 1, \cdots, n-1$$

$$P(Y=n/X) = \frac{1}{e^{\left[\ln \frac{\delta_0^{[m]}}{\delta_0^{[M]}} \times \sum_i X_i \omega_1 \right]} + e^{\left[\ln \frac{\delta_1^{[m]}}{\delta_1^{[M]}} \times \sum_i X_i \omega_2 \right] + \cdots + 1}} \tag{19}$$

Where $\omega_0, \omega_1, \cdots, \omega_{n-1}$ are the weights.

5. Parameter Estimation for the Bayesian Multinomial Regression using Gaussian Naïve Bayes

The log of the conditional likelihood:

$$\ln P\left(Y_i/X_i\, W\right) = \sum_{i=1}^{n} Y_i \ln P\left(Y_i = 0/X_i\right) + \left(1 - Y_{i-0,2,\cdots,n}\right) \ln P\left(Y_{i-0,2,\cdots,n} = 1/X_i\right) +$$

$$\cdots + \left(1 - Y_{i-0,2,\cdots,n-1}\right) \ln P\left(Y_{i-0,2,\cdots,n-1} = 1/X_i\right) \quad i = 0, 1, 2, \cdots, n \tag{20}$$

$$= \sum_{i=1}^{n} Y_i \ln \delta_0^{[M]} + \left(1 - Y_{i-0,2,\cdots,n}\right) \ln \left(1 - \delta_0^{[M]}\right) + \cdots + \left(1 - Y_{i-0,2,\cdots,n-1}\right) \ln \left(1 - \sum_{i=1}^{M-1} \delta_i^{[m]}\right) \tag{21}$$

$$= \sum_{i=1}^{n} Y_i \ln \delta_0^{[M]} + (1 - Y_{i-0,2,\cdots,n}) \ln \left(1 - \delta_0^{[M]}\right) + \cdots + (1 - Y_{i-0,2,\cdots,n-1}) \ln \sum_{i=1}^{M-1} \delta_i^{[m]} \tag{22}$$

$$= \sum_{i=1}^{n} (Y_i + (1 - Y_{i-0,2,\cdots,n}) + \cdots + (1 - Y_{i-0,2,\cdots,n-1})) \ln \left(\delta_0^{[M]} \left(1 - \delta_0^{[M]}\right) + \cdots + \sum_{i=1}^{M-1} \delta_i^{[m]}\right) \tag{23}$$

This can be compared to the canonical form of Generalized Linear Model (GLM) $\sum_{i=1}^{n} \left[\frac{x_i \delta_i^{[m]}(\omega) - b(\delta_i^{[m]}(\omega))}{a_i(\phi)} + c(y_i, \delta^{[m]})\right]$, then solve via iterative optimization of Reweighted Iterative Least Square procedure, gradient ascent or Newton Rapshon method.

6. Non-parametric Classification

Non-parametric uses bandwidth say $h > 0$ near a point say x_0. Assuming the Probability Density Function for the observations is $\widehat{f}(x_0)$ with corresponding Cumulative Density Function (CDF) say $\widehat{F}(x_0)$. The derivative at the point x_0 is,

$$\widehat{f}(x_0) = \frac{F\left(x_0 + \frac{h}{2}\right) - F\left(x_0 - \frac{h}{2}\right)}{h} \tag{24}$$

The numerator can be written as $P\left(x_0 - \frac{h}{2} \leq x \leq x_0 + \frac{h}{2}\right) = F\left(x_0 + \frac{h}{2}\right) - F\left(x_0 - \frac{h}{2}\right) = \frac{x_i \in N(x_o)}{n}$

So,

$$\widehat{f}(x_0) = \frac{x_i \in N(x_o)}{nh} \tag{25}$$

Then non-parametric Kernel Density Estimation (KDE) via Naïve Bayes: suppose in a class of C_k with ith features, such that in each feature of $(x_1, x_2, \cdots, x_n)$. The conditional probability $f_i(x_i|C = c)$, the probability that feature value in the ith position x_i given class c. The training data for the KDE (X, C). Its kernel density estimator can be drawn as follows:

$$\widehat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{26}$$

Where $h > 0$ is the smoothing parameter called the bandwidth (otherwise known as the smoothing parameter of the KDE estimator $\widehat{f}$), n is the sample size, and $K(\cdot)$ is the kernel function with the properties that $K(\cdot) : \Re \to \Re$, such that $\int_{-\infty}^{\infty} K(\cdot) = 1$, $E(\cdot) = 0$. The kernels to be used are Gaussian and biweight.

The Gaussian based Kernel Density Naïve Bayes probability is

$$f_i(x_i|C = c) = \frac{1}{N_c h} \sum_{j=1}^{N_c} K\left(x_i, x_{j|i|c}\right) \; K(\delta_i^{[m]}, \sigma^2) = \frac{1}{\sqrt{2\pi}} e^{\left(\frac{\left(x_i - \delta_i^{[m]}\right)}{2\sigma_i^2}\right)^2} \tag{27}$$

Where the function K(.) is the Gaussian function kernel with mean zero and variance 1. N_c is the number of observations X belonging to class c, $x_{j|i|c}$. The value in the ith position of the jth input $(x_1, x_2, \cdots, x_n)$ in class c, such "h" is a bandwidth, or a smoothing parameter.

The Bi-weight based Kernel Density Naïve Bayes probability is

$$f_i\left(x_i|C=c\right)=\frac{1}{N_c h}\sum_{j=1}^{N_c}K\left(x_i,x_{j|i|c}\right)\ K(\delta_i^{[m]},\sigma^2)=\frac{15\left(1-\left(X_i-\delta_i^{[m]}\right)^2\right)^2}{16}I\left(|\frac{X_i-\delta_i^{[m]}}{h}|\leq 1\right) \tag{28}$$

However, the conditional probabilities can also be estimated using Laplace Smoothing (LS) via

$$f_i\left(x_i|C=c\right)=\frac{m_{y_i|c}+1}{n_c+1} \tag{29}$$

Such that $m_{x_i|c}$ is the number of input X pertaining to class c and i^{th} position as same x_i. n_c is the total number of observation X in class c. k is the number of possible unique feature values for input X. According to Silverman (1986), the rule-of-thumb method for estimating bandwidth "h" is

$$\widehat{h}=\left(\frac{4\widehat{\sigma}^5}{3n}\right)^{1/5}\approx 1.06\widehat{\sigma}n^{-1/5} \tag{30}$$

Where σ^2 the sample standard deviation and n is the number of samples. Silverman's rule of thumb assumes that the underlying kernel is Gaussian.

7. Numerical Analysis

The dataset to be used to validate the Naïve Bayes classifiers contains information on 78 people using one of three diets (Diet A, B, or C). The dataset was extracted from the website of Department of mathematics and statistics, University of Sheffield (https://sheffield.ac.uk/mash/statistics/datasets), the dataset can also be found on stcp-dataset-diet_des. The dataset contains variables like participants number, binary variables - gender and three different types of diets taken. Additionally, the dataset consists of scale variables - Age (in years), Height (in cm), weight (in kg) before the diet (that is, pre-weight), weight (in kg) gained after 6 weeks of diet. Ellen Marshal, University of Sheffield, collated the data. The research questions were meant for studying the influence of the three different types of diets taken on the variables mentioned (with the exception of the participants' number and gender).

Table 1. Posterior Bayes Naïve: Diets' Prior and A Prior Probabilities with Variables' Means and Standard Errors

Variable	Intercept	Age	Height	Pre.weight	weight after six weeks	Prior Probabilities	A Prior Probabilities
Diet A	1.2972 (0.1597)	40.8750 (9.72810)	170.2917 (10.9484)	72.8750 (8.3838)	69.5750 (8.3984)	0.3077	0.3077
Diet B	3.2172 (0.0467)	39.0000 (9.5111)	174.8518 (12.0821)	71.1111 (10.0932)	68.0852 (10.2172)	0.3462	0.3462
Diet C	0.2781 (0.3610)	37.7778 (9.3155)	167.2593 (9.7096)	73.6296 (7.6064)	68.4815 (8.2428)	0.3462	0.3462

Keys: x=Means; y=Median; bw=Bandwidths

Source: Authors' Computation (2022).

From table 1 above, diet type (A, B and C) as the dependent variable (subdivided into three categorical), such that age, height, pre-weight and weight after six weeks are the independent variables to be evaluated. The positive signs of means of 40.8750, 170.2917, 72.8750, 69.5750 for age, height, pre-weight and weight after six weeks variables respectively imply that there is a positive contribution or relationship to Diet A. In a similar vein, positive signs of means of 39.0000, 174.8518, 71.1111, 68.0852 for age, height, pre-weight and weight after six weeks variables respectively imply that there is a positive contribution or relationship to Diet B. While positive signs of means of 37.7778, 167.2593, 73.6296, 68.4815 for age, height, pre-weight and weight after six weeks variables respectively imply that, there is a positive contribution or relationship to Diet C. It is to be noted that Diet C is the most significant with the mentioned contributors due its bareness minimum standard errors of 9.3155, 9.7096, 7.6064 and 8.2428 respectively for age, height, pre-weight and weight after six weeks variables.

$$P(Y = Diet\,A/X) = \frac{1}{1 + \exp\left(1.2972 + 40.8750X_1 + 170.2917X_2 + 72.8750X_3 + 69.5750X_4\right)}$$

$$P(Y = Diet\,B/X) = \frac{1}{1 + \exp\left(3.2172 + 39X_1 + 174.8518X_2 + 71.1111X_3 + 68.0852X_4\right)}$$

Then,

$$P(Y = Diet\,C/X) = \frac{1}{1 + \exp\left(0.2181 + 37.7778X_1 + 167.2593X_2 + 73.6296X_3 + 68.4815X_4\right)}$$

It is to be noted that X_1=Age, X_2=Height, X_3=Pre.weight, X_4=Weight after six weeks.

Table 2. Predictive Probabilities

Variables	Age	Height	Pre.weight	Weight after weeks
Multinomial Sigmoid $\sigma(a_i)$	0.4107	0.3673	0.2219	0.0024

Source: Authors' Computation (2022).

From table 3 above, it can be deduced that "age" is the most contributing factor among the four covariates to diet, followed "height", "pre-weight" and "weight after six weeks".

Table 3. Non-parametrically Gaussian KDE: Diets' Prior and A prior Probabilities with Variables' Means and Median

Variables	Age			Height			Pre. Weight			Weight after Weeks		
	x	y	bw	x	y	bw	x	y	bw	x	y	bw
Diet A	41.000	0.2090	4.446	178.5	0.2123	3.112	73.00	0.250	3.996	69.25	0.2505	3.913
	(41.000)	(0.1853)		(178.5)	(0.3715)		(73.00)	(0.2463)		(69.25)	(0.2306)	
Diet B	35.000	0.1398	4.169	179.5	0.1896	5.385	179.5	0.1896	5.385	79.00	0.1820	4.378
	(35.000)	(0.1398)		(179.5)	(0.1575)		(179.5)	(0.1575)		(79.00)	(0.1752)	
Diet C	39.000	0.1897	4.802	162	0.1986	4.343	162	0.1987	4.343	67.45	0.2604	3.837
	(39.000)	(0.1897)		(162)	(0.4673)		(162)	(0.4673)		(67.45)	(0.2804)	

Source: Authors' Computation (2022).

Table 4. Non-parametrically Bi-weight KDE: Diets' Prior and A prior Probabilities with Variables' Means and Median

Variables	Age			Height			Pre. Weight			Weight after Weeks		
	x	y	bw	x	y	bw	x	y	bw	x	y	bw
Diet A	41.000	0.0154	4.446	178.5	0.0157	3.112	73.00	0.0181	3.996	69.25	0.0185	3.913
	(41.000)	(0.0139)		(178.5)	(0.0077)		(73.00)	(0.0185)		(69.25)	(0.0174)	
Diet B	35.000	0.0159	4.169	179.5	0.0118	5.385	80.5	0.0050	4.699	79.00	0.0134	4.378
	(35.000)	(0.0113)		(179.5)	(0.0140)		(80.5)	(0.0137)		(79.00)	(0.0036)	
Diet C	39.000	0.0150	4.802	162	0.0147	4.343	74	0.0205	3.474	67.45	0.0192	3.837
	(39.000)	(0.0139)		(162)	(0.0100)		(74)	(0.0189)		(67.45)	(0.0207)	

Source: Authors' Computation (2022).

Table 5. Non-parametrically Silverman's Rule of Thumb Bandwidth Selector: Diets' Prior and A prior Probabilities with Variables' Means and Median

Variables	Age			Height			Pre. Weight			Weight after Weeks		
	x	y	bw	x	y	bw	x	y	bw	x	y	bw
Diet A	41.000	0.1809	6.121	178.5	0.0068	4.142	73.00	0.0181	3.996	69.25	0.2546	3.768
	(41.000)	(0.1365)		(178.5)	(0.0143)		(73.00)	(0.0185)		(69.25)	(0.2373)	
Diet B	35.000	0.1959	5.162	179.5	0.2119	4.131	80.5	0.0050	4.699	79.00	0.2034	3.074
	(35.000)	(0.4589)		(179.5)	(0.1674)		(80.5)	(0.0137)		(79.00)	(0.2382)	
Diet C	39.000	0.0123	5.536	162	0.1815	5.414	74	0.0205	3.474	67.45	0.2324	4.878
	(39.000)	(0.0140)		(162)	(0.3862)		(74)	(0.0189)		(67.45)	(0.2142)	

Source: Authors' Computation (2022).

Table 6. Non-parametrically Adjusted Bandwidth (1.5): Diets' Prior and A prior Probabilities with Variables' Means and Median

Variables	Age			Height			Pre. Weight			Weight after Weeks		
	x	y	bw	x	y	bw	x	y	bw	x	y	bw
Diet A	41.000	0.1732	6.67	178.5	0.1851	4.669	73.00	0.2048	5.994	69.25	0.2546	5.869
	(41.000)	(0.4605)		(178.5)	(0.3300)		(73.00)	(0.1610)		(69.25)	(0.1579)	
Diet B	35.000	0.1790	6.254	179.5	0.1545	8.078	80.5	0.1548	7.048	79.00	0.1545	8.078
	(35.000)	(0.4157)		(179.5)	(0.1360)		(80.5)	(0.2164)		(79.00)	(0.1360)	
Diet C	39.000	0.1664	7.204	162	0.1667	6.514	74	0.2280	5.211	67.45	0.2131	5.756
	(39.000)	(0.4683)		(162)	(0.3398)		(74)	(0.1778)		(67.45)	(0.1770)	

Source: Authors' Computation (2022).

The data sample is "x" also termed as the mean, the points of the grid at which the density derivative is to be estimated is "y", also termed as the median and "h" the smoothing bandwidth via unbiased cross validation. It is to be noted from table 3 to table 6 for non-parametrically Gaussian KDE, bi-weight KDE, Silverman's Rule of Thumb bandwidth Selector, and Adjusted Bandwidth (1.5) respectively produced the same prior probabilities of 0.3077, 0.3462, and 0.3462; and A prior probabilities of 0.3077, 0.3462, and 0.3462 for Diet A, Diet B, and Diet C at different smoothing bandwidths, the same set of priors with that of Posterior Bayes Naïve. Gaussian KDE and bi-weight KDE produced the bareness minimum bandwidths across the three Diets for age, height, pre-weight, weight after six weeks in comparison to Silverman's Rule of Thumb Bandwidth Selector and adjusted bandwidth (1.5).

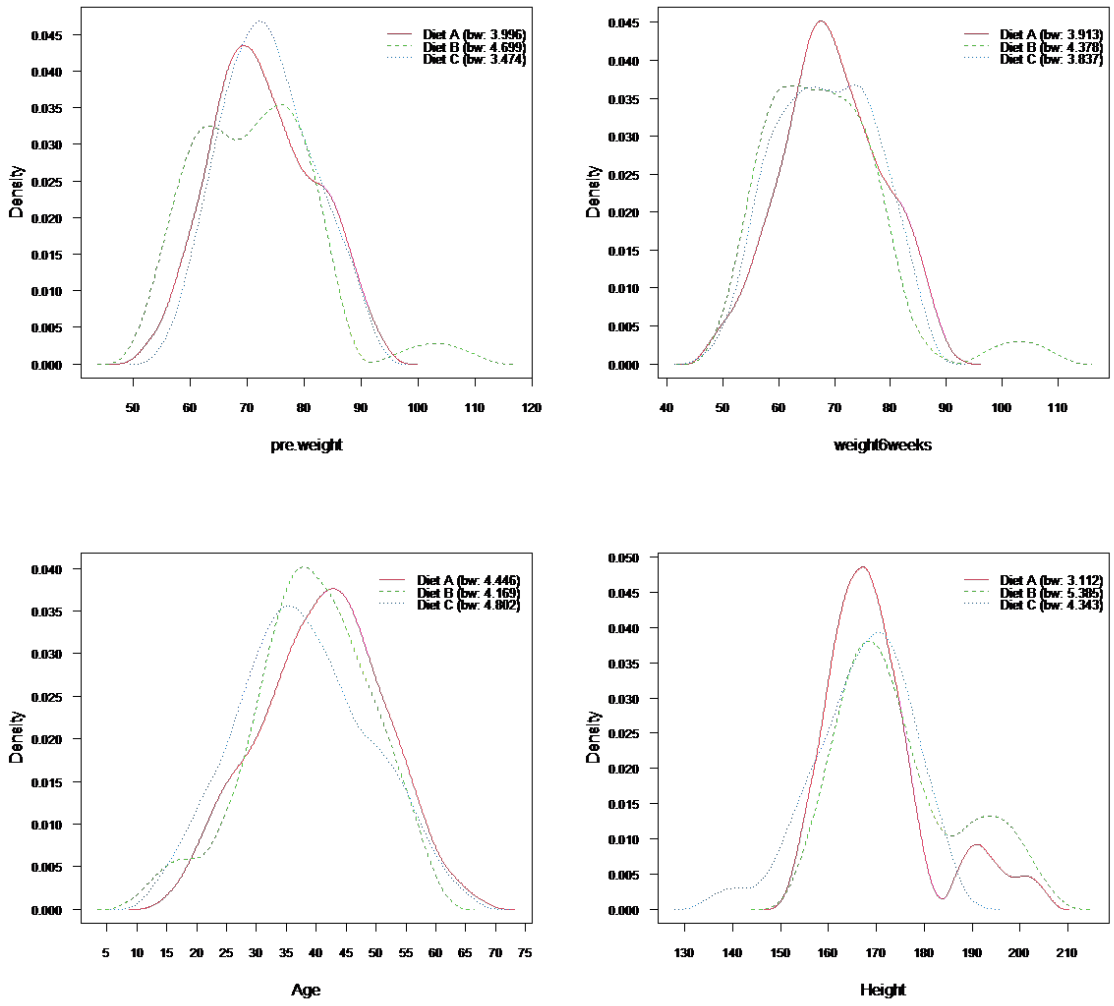

FIGURE 1. Visualization of the Naïve Bayes Conditional Density Plots

Source:Authors' Computation (2022).

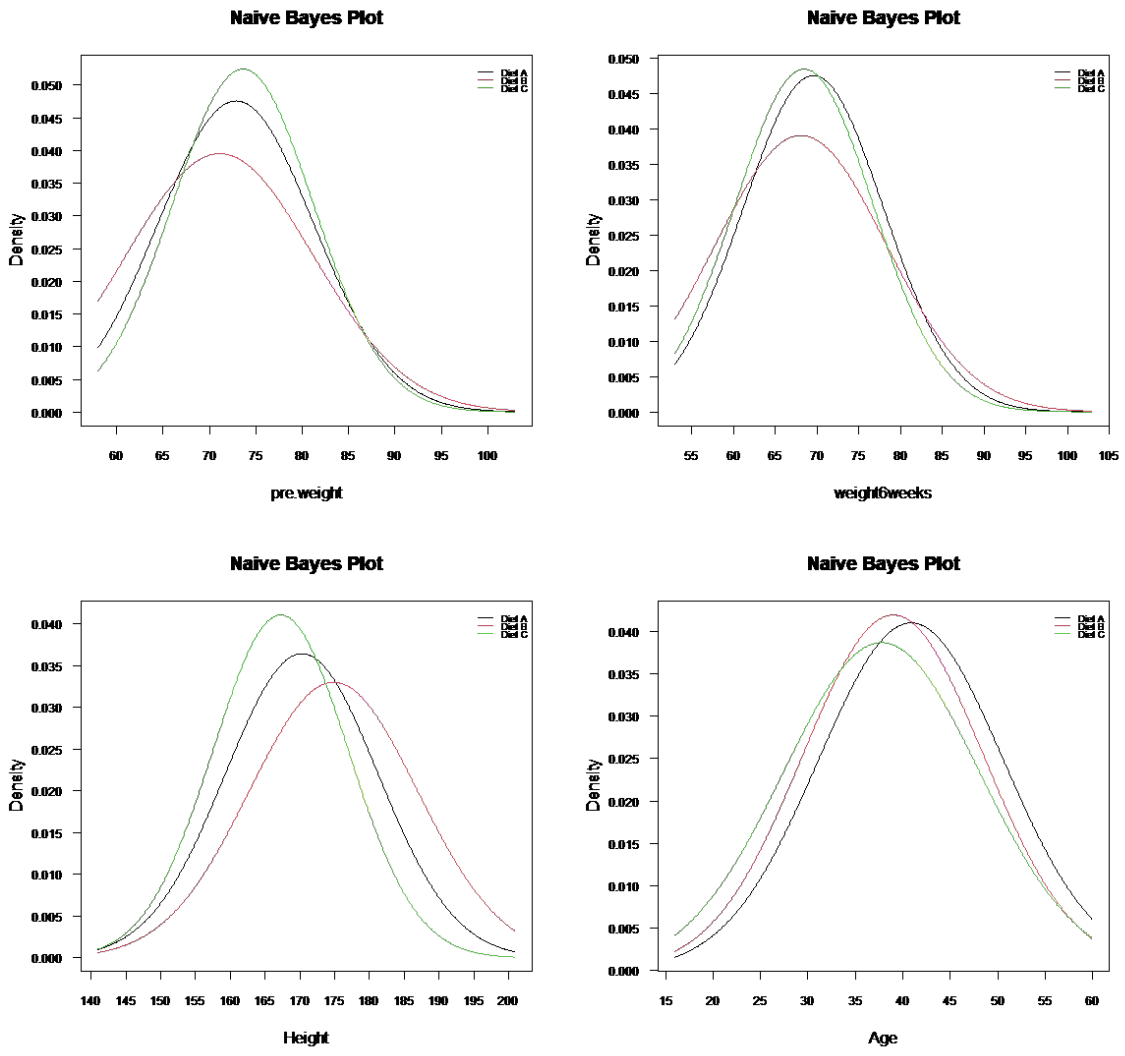

FIGURE 2. Visualization of the Kernel Conditional Densities of the Corresponds

Source:Authors' Computation (2022).

8. Conclusions

This paper introduced a Bayesian and nonparametric framework for solving and analyzing multinomial regression. The framework adopted a Naïve Bayes classifier for the Bayesian approach and nonparametric approach via kernel density estimation. The kernels adopted in this research are Gaussian KDE, bi-weight KDE, Silverman's Rule of Thumb bandwidth Selector, and Adjusted Bandwidth (1.5). In conclusion, the Gaussian KDE and bi-weight KDE produced the bareness minimum bandwidths across the three categorical responses for the four influencers. Both the Naïve Bayes classifier and nonparametric KDE for the multinomial regression produced the same prior probabilities of 0.3077, 0.3462, and 0.3462; and A prior probabilities of 0.3077, 0.3462, and 0.3462 for Diet A, Diet B, and Diet C at different smoothing bandwidths. An extension of this work can be carried-out by extending the Naïve Bayes classifier and non-parametric KDE to Negative-Binomial or zero-inflated regression for modeling count responses, that are usually affected by over-dispersion of count outcomes.

References

[1] D. Andrea, O. Nicola, Approximate Bayesian logistic regression via penalized likelihood estimation with data augmentation, Unit of Biostatistics and Unit of Nutritional Epidemiology Institute of Environmental Medicine, Karolinska. (2014) 1-24.

[2] L. Blizzard, D.W. Hosmer, The log multinomial regression model for nominal outcomes with more than two attributes, Biom. J. 49 (2007) 889–902. https://doi.org/10.1002/bimj.200610377.

[3] H. Chen, S. Hu, R. Hua, X. Zhao, Improved naive Bayes classification algorithm for traffic risk management, EURASIP J. Adv. Signal Process. 2021 (2021) 30. https://doi.org/10.1186/s13634-021-00742-6.

[4] H. Chen, D. Fu, An improved naïve bayes classifier for large scale text, Advances in Intelligent Systems Research 146 (2018). 2^{nd} International Conference on Artificial Intelligence: Technologies and Applications (ICAITA 2018).

[5] Y. Croissant, mlogit: Multinomial Logit Models. R package version 1.0-3.1 (2020). https://CRAN.R-project.org/package=mlogit.

[6] A. Hasan, Z. Wang, A.S. Mahani, Fast estimation of multinomial logit models: R Package mnlogit, J. Stat. Soft. 75 (2016) 1-24. https://doi.org/10.18637/jss.v075.i03.

[7] J.E. Johndrow, A. Smith, N. Pillai, D.B. Dunson, MCMC for imbalanced categorical data, J. Amer. Stat. Assoc. 114 (2019) 1394–1403. https://doi.org/10.1080/01621459.2018.1505626.

[8] P. Kaviani, D. Sunita, Short survey on Naïve Bayes algorithm, Int. J. Adv. Eng. Res. Develop. 4 (2017) 2348-4470.

[9] A. Kelly, M.A. Johnson, Investigating the statistical assumptions of Naïve Bayes classifiers, in: 2021 55^{th} Annual Conference on Information Sciences and Systems (CISS), IEEE, Baltimore, MD, USA, 2021: pp. 1–6. https://doi.org/10.1109/CISS50987.2021.9400215.

[10] M.E. Glickman, D.A. Dyk, Basic Bayesian methods, in: W.T. Ambrosius (Ed.), Topics in Biostatistics, Humana Press, Totowa, NJ, 2007: pp. 319–338. https://doi.org/10.1007/978-1-59745-530-5_16.

[11] B. Nandram, A Bayesian approach to linking a survey and a census via small areas, Stats. 4 (2021) 509–528. https://doi.org/10.3390/stats4020031.

[12] B. Nandram, L. Chen, S. Fu, B. Manandhar, Bayesian logistic regression for small areas with numerous households, Stat. Appl. 16 (2018) 171–205.

[13] N. Salmi, Z. Rustam, Naïve Bayes classifier models for predicting the colon cancer, IOP Conf. Ser.: Mater. Sci. Eng. 546 (2019) 052068. https://doi.org/10.1088/1757-899X/546/5/052068.

[14] M. Sarrias, R. Daziano, Multinomial logit models with continuous and discrete individual heterogeneity in R: The gmnl Package, J. Stat. Soft. 79 (2017). https://doi.org/10.18637/jss.v079.i02.

[15] B.W. Silverman, Density estimation for statistics and data analysis, Chapman and Hall, New York (1986).

[16] S. Sinharay, Discrete probability distributions, In International Encyclopedia of Education (3^{rd} Edition) (2010).

[17] R.J. Poovaraghan, M.V.K. Priya, S.S.Vamsi, M. Mewara, S. Loganathan, Fake news accuracy using Naïve Bayes classifier, Int. J. Recent Technol. Eng. 8 (2019) 45-78.

[18] A.P. Wibawa, A.C. Kurniawan, D.M.P. Murti, et al. Naïve Bayes classifier for journal quartile classification, Int. J. Recent Contrib. Eng. Sci. IT. 7 (2019) 91. https://doi.org/10.3991/ijes.v7i2.10659.

[19] G. Wioletta, The advantages of Bayesian methods over classical methods in the context of credible intervals, Inform. Syst. Manage. 4 (2015) 53-63.

[20] S. Xu, Bayesian Naïve Bayes classifiers to text classification, J. Inform. Sci. 44 (2018) 48–59. https://doi.org/10.1177/0165551516677946.

[21] W. Zhang, F. Gao, An improvement to Naive Bayes for text classification, Procedia Eng. 15 (2011) 2160–2164. https://doi.org/10.1016/j.proeng.2011.08.404.