Gender-Based Comparative Study of Type 2 Diabetes Risk Factors in Kolkata, India: Frequentist versus Bayesian Approach in Machine Learning

Rahul Jain[1], *Anoushka* Saha[2], *Durba* Bhattacharya[2], *Madhura* Das Gupta[2], *Sourav* Chowdhury[3,*], *Gourav* Daga[2], *Suparna* Roychowdhury[3]

[1]Belle Vue Clinic, Kolkata, India
[2]Department of Statistics, St. Xavier's College (Autonomous), Kolkata, India
[3]Postgraduate and Research Department of Physics, St. Xavier's College (Autonomous), Kolkata, India

Abstract. Type 2 diabetes mellitus represents a prevalent and widespread global health concern, necessitating a comprehensive assessment of its risk factors. This study aims to evaluate and compare the predictive performance of frequentist and Bayesian machine learning models in assessing the risk of Type 2 diabetes mellitus based on age, lifestyle, BMI, and waist-to-height ratio among males and females in Kolkata, West Bengal, India. The analysis utilizes data from patients observed in the outpatient consultation department of Belle Vue Clinic in Kolkata. The frequentist models employed include Random Forest (RF), and Support Vector Classifier (SVC), while their Bayesian counterparts - Bayesian Additive Regression Trees (BART), and Relevance Vector Machine (RVM) were also used. Our findings indicate that for males, BMI is the most important predictor of Type 2 Diabetes, whereas for females, Whtr is identified as the most important predictor. This study highlights gender-specific differences in risk factors for Type 2 diabetes mellitus and contributes to understanding the effectiveness of various modeling approaches in predicting risk within this population. The insights gained from this research can inform more targeted healthcare interventions and public health strategies.

1 Introduction

Type 2 diabetes is a rapidly growing non-communicable disease in India, ranking second worldwide with over 74 million cases in 2021. The IDF Diabetes Atlas (2021) [1] projects a 68% increase to 124.9 million individuals by 2045, posing a substantial health challenge.

Common Type 2 diabetes risk factors include excess weight, obesity, and sedentary lifestyle, but other factors still remain largely unexplored [2-4]. Existing anthropometric measures like BMI and waist circumference have limitations in assessing body fat distribution and accounting for ethnic variations [5]. Research suggests that waist-to-height ratio (Whtr), a measure of central obesity, may provide better insights into predicting cardio-metabolic abnormalities compared to BMI and waist circumference [6]. Global studies demonstrate variations in the performance of anthropometric measures for predicting Type 2 Diabetes Mellitus risk across diverse subpopulations [7]. Therefore, it is crucial to assess the predictive capabilities of these measures within distinct ethnic or geographic groups.

* Corresponding author: chowdhury95sourav@gmail.com

Moreover, the prevalence of diabetes differs between obese men and women [7], with age consistently recognized as a significant risk factor for Type 2 diabetes [8].

This study is aimed towards exploring the application of machine learning and Bayesian models in healthcare, specifically within the context of predicting diabetes risk in Kolkata, West Bengal, India. This work also assesses and compares the predictive capabilities of key variables: BMI, WHR, age, and lifestyle, separately for males and females using appropriate machine learning and Bayesian algorithms. The analysis incorporates various machine learning algorithms such as the Random Forest Classifier (RF) and the Support Vector Classifier (SVC). The Bayesian models employed include Bayesian Additive Regression Trees (BART) and Relevance Vector Machine (RVM). The predictive outcomes of these models are depicted through confusion matrices, enabling the calculation of metrics such as accuracy, precision, recall, and F1 score.

The rest of the paper is structured as follows: Section 2 presents a comprehensive description of the data and its collection procedure, including the definition of the variables under study. In section 3, a short summary of the data is presented. Section 4 and Section 5 respectively describe the machine learning and Bayesian models built on the data. Finally, in Section 6, the work is summarized, and concluding remarks are made.

2 Methods

A cross-sectional study was performed in the outpatient department of Belle Vue Clinic from March to May 2022, involving 428 patients. Among them, 211 were diagnosed with Type 2 diabetes mellitus, while 217 were tested negative. Following the guidelines of the Ethics Committee of the hospital, the participants were informed about the purpose of the study before collecting data.

Diabetes screening involved fasting plasma glucose measurement using the Hexokinase method after an overnight fast, with a cutoff of $\geq$ 126 mg/dl indicating diabetes in the fasting state. 2-hour postprandial blood glucose of $\geq$ 200mg/dl and/or HbA1c of $\geq$ 6.5 (measured by HPLC method) or above were taken as Diabetic.

Height was measured with the subject standing against a wall-mounted scale, weight with a digital weighing machine while standing, and waist circumference at the midpoint of the Iliac crest anteriorly and the lower ribs with a flexible measuring tape, all while the subject stood and looked forward. Each measurement was taken twice, and the averages were recorded.

Table 1. Description of the variables under study

Predictors	**Type**	**Observed Range / Categories**
Age	Continuous	Min: 20, Max: 88
Sex	Categorical	204 Males and 224 Females
Height	Continuous	Min:139 cm, Max: 189.5 cm
Weight	Continuous	Min: 41.8 kg, Max:144.4 kg
Waist Circumference	Continuous	Min=64; Max=143
Lifestyle	Categorical	3 categories: Does not exercise, Moderately exercises, Exercises daily.

Information on Lifestyle activity and Gender was supplied by the respondents themselves. BMI and Waist-to-Height Ratio (henceforth referred to as Whtr) were derived from the

variables described in Table 1. Considering the increased risk of diabetes with age, particularly at age ≥ 45, as established by previous studies [8] and the National Institute of Diabetes and Digestive and Kidney Diseases, the data was stratified by age and gender.

The interrelationship between various anthropometric variables was studied using correlation heat maps [11]. The dataset was divided randomly into 80% training and 20% test sets. Machine learning models random forest, Support Vector Classifier models, and Bayesian models like Relevance Vector Machine and Bayesian Additive Regression Trees were applied separately for male and female Type 2 diabetes risk prediction. Furthermore, performance metrics like accuracy, precision, recall, and F1 score were calculated for comparison of machine learning and Bayesian models. Python and R software were used to perform the statistical analysis and machine learning.

3 Data Summary

The mean age of the participants was 53.3 years, with a standard deviation of 14.7 years. The mean height of the participants was 162.1 cm with a standard deviation of 9.6 cm. The mean weight of the participants was 72.7 kg with a standard deviation of 15.4 kg.

Fig. 1. Correlation heat map for all the possible predictor variables: Age, Height, Weight, BMI, Waist circumference, and Whtr in our study.

The system has multicollinearity between weight, BMI, and waist circumference, which is understood from the correlation plot (Fig. 1). Thus, we have excluded weight and waist circumference from the list of the predictor variables during the learning process. Height is not directly correlated with diabetes, thus, height is also excluded from the list. Also, there is a high correlation (r = 0.82) between BMI and Whtr. However, BMI and Whtr both are important predictor variables for the classification of diabetes and thus we have retained these two variables as predictors [4, 5, 6]. Also, for fixed levels of height or weight, on an average, the females tend to have more BMI than the males. For fixed levels of weight, on an average, the females tend to have a higher waist-to-height ratio (Whtr) [11].

Thus, there are four predictor variables in our machine learning study: Age, BMI, Lifestyle, and Whtr. Although age is a continuous variable, we have converted age to a categorical variable in our study. It was already mentioned in the previous section, the risk of diabetes increases with age as found in earlier studies [8]. Thus, age is encoded into two

categories: age ≥ 45 and age < 45. The implementation of different machine learning algorithms with these four predictor variables is discussed in the next section.

4 Machine Learning Models

In this section, we have applied frequently used machine learning models: Random Forest Classifier (RF), and Support Vector Classifier (SVC) to our data. These models are separately implemented for males and females. Also, the four predictor variables used here are age (encoded), BMI, lifestyle, and Whtr.

4.1 Random Forest (RF)

Random forest models were fitted, separately for males and females using R, considering the predictors Age, Lifestyle, BMI, and Whtr. We have taken both BMI and Whtr in the same model because the random forest algorithm is not much affected by the presence of multicollinearity. The number of trees constructed was 200 and the number of variables considered at each split was 2. The out-of-bag estimate of error rate is 35.91% for females and 30.54% for males.

Table 2. Confusion matrix and performance measures for the Random Forest model based on Age, BMI, Whtr, and Lifestyle.

Females			**Males**		
	Predicted			**Predicted**	
Observed	**0**	**1**	**Observed**	**0**	**1**
0	13	12	**0**	11	6
1	3	15	**1**	8	12
Accuracy = 0.651			Accuracy = 0.622		
Recall = 0.833			Recall = 0.6		
Precision = 0.556			Precision = 0.667		
F1 Score = 0.667			F1 Score = 0.632		

The confusion matrices obtained from predictions on the test set were used to evaluate the performance measures for the different models. Table 2 enlists the confusion matrices for males and females for RF.

Table 3. Variable Importance Measure (Mean Decrease in Gini in index) for Random Forest model based on Age, Lifestyle, BMI, and Whtr for Females and Males.

Predictor	**Mean Decrease in Gini Index**	
	Female	**Male**
Age	14.417	16.163
Lifestyle	5.165	6.483
BMI	29.507	27.150
Whtr	31.883	26.342

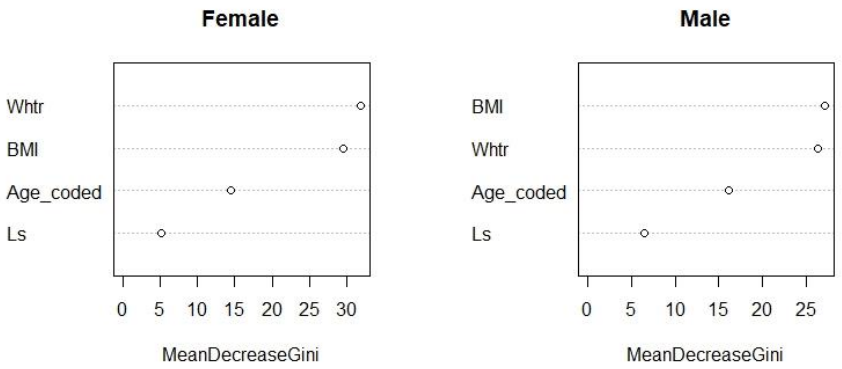

Fig. 2. Variable Importance Plot for the Random Forest Model based on Age, BMI, Whtr, and Lifestyle for Females and Males.

Table 3 and Fig. 2 obtained from the random forest models show that for females, the mean decrease in the Gini index is the maximum for Whtr followed by that of BMI. This implies that the total decrease in the node impurity that results from a split over Whtr is the maximum as compared to splits over the other variables. For males, the mean decrease in Gini index is the maximum for BMI, followed by that of Whtr. Hence, we can conclude that Whtr is the most important predictor for Type 2 Diabetes Mellitus in females and BMI is the most important predictor for Type 2 Diabetes Mellitus in males. This conclusion is at par with the one drawn from the Logistic Regression models [11].

4.2 Support Vector Classifier (SVC)

The Support Vector Classifier algorithm is also not affected by the presence of multicollinearity. Separate models for males and females were fitted (using Python), considering the predictors: Age, Lifestyle, BMI, and Whtr. The minimum error was obtained for cost C=0.1 and with the linear kernel using 10-fold cross-validation. The confusion matrices (Table 4) of SVC are used here to evaluate the performance based on the test data.

Table 4. Confusion matrix and performance measures for Support vector machine classifier model based on Age, BMI, Whtr, and Lifestyle

Females			**Males**		
	Predicted			**Predicted**	
Observed	**0**	**1**	**Observed**	**0**	**1**
0	10	15	**0**	9	8
1	2	16	**1**	3	17
Accuracy = 0.605 Recall = 0.889 Precision = 0.516 F1 Score = 0.653			Accuracy = 0.703 Recall = 0.85 Precision = 0.68 F1 Score = 0.756		

5 Bayesian Models

In this section, we have applied the Bayesian counterpart of the frequently used machine learning models (RF and SVC): Bayesian Additive Regression Trees (BART) and Relevance Vector Machine (RVM) to our data. These models are separately implemented for males and females as previously. Also, the four predictor variables used here are age (encoded), BMI, lifestyle, and Whtr.

5.1 Bayesian Additive Regression Tree (BART)

Bayesian Additive Regression Trees were built on the training data separately for males and females [10]. The number of trees to be grown in the sum-of-trees model is chosen to be 200.

Table 5. Confusion matrix and performance measures for BART model based on Age, BMI, Whtr, and Lifestyle

Females			**Males**		
	Predicted			**Predicted**	
Observed	**0**	**1**	**Observed**	**0**	**1**
0	15	10	**0**	11	6
1	6	12	**1**	7	13
Accuracy = 0.628			Accuracy = 0.649		
Recall = 0.667			Recall = 0.65		
Precision = 0.545			Precision = 0.684		
F1 Score = 0.6			F1 Score = 0.667		

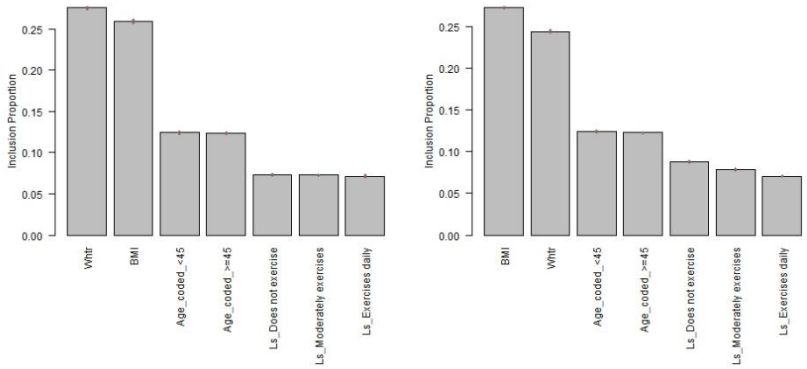

Fig. 3. Left Panel **(a):** Variable Inclusion Proportions Plot for BART Model based on Age, BMI, Whtr, and Lifestyle for Females. Right Panel **(b):** Variable Inclusion Proportions Plot for BART Model based on Age, BMI, Whtr, and Lifestyle for Males.

The number of MCMC samples to be discarded as "burn-in" was taken to be 10000 and the number of MCMC samples to be drawn from the posterior distribution is taken to be 30000. The confusion matrices (Table 5) obtained from predictions on the test set were used to evaluate the performance of BART.

The variable inclusion proportions obtained from the BART model for females in Fig. 3(a) show that the relative influence of Whtr is the highest followed by that of BMI. On the other hand, the variable inclusion proportions obtained from the BART model for males in Fig. 3(b) show that the relative influence of BMI is the highest followed by that of Whtr. Hence, we can say that Whtr is the most important predictor for females and BMI is the most important predictor for males. This conclusion is at par with the one drawn from the Random Forest models.

5.2 Relevance Vector Machine (RVM)

The Relevance Vector Machine (RVM) algorithm is a Bayesian counterpart of the Support Vector Classifier (SVC). Separate models of RVM for males and females were fitted (using Python), considering the predictors Age, Lifestyle, BMI, and Whtr. Similar to the SVC, the cost for RVM is chosen as C=0.1 with a linear kernel. The confusion matrices (Table 6) are used here to evaluate the performance based on the test data of RVM.

Table 6. Confusion matrix and performance measures for Relevance Vector Machine model based on Age, BMI, Whtr, and Lifestyle

Females			**Males**		
	Predicted			**Predicted**	
Observed	**0**	**1**	**Observed**	**0**	**1**
0	10	15	**0**	0	17
1	2	16	**1**	0	20
Accuracy = 0.605			Accuracy = 0.541		
Recall = 0.889			Recall = 1.0		
Precision = 0.516			Precision = 0.541		
F1 Score = 0.653			F1 Score = 0.701		

6 Discussion and Conclusion

In conclusion, our gender-based comparative analysis of machine learning and Bayesian algorithms revealed that the Bayesian Additive Regression Tree (BART) has performed better than other frequently used models (RF and SVC), demonstrating better accuracy and F1 scores. This makes BART the preferred model for diabetes risk prediction in our study. RVM could not perform well in this data while the Support Vector Classifier excelled in predicting risk for males but underperformed for females. The Random Forest model, on the other hand, delivered better results for females. These findings not only underscore the effectiveness of BART in this context but also highlight the importance of gender-specific model selection in diabetes risk assessment. The observed discrepancies between model performances across genders suggest the need for tailored approaches in predictive healthcare analytics.

One of the key findings in our study is that BMI does not significantly predict diabetes risk in females, unlike in males. This finding contrasts with some previous studies [2-4] but aligns with the conclusions of [11]. Additionally, we identify waist-to-height ratio (Whtr) and age ≥ 45 as significant risk factors for Type 2 diabetes in both genders in Kolkata, emphasizing the need for targeted preventive measures for individuals in this age group. Notably, Whtr emerges as a superior predictor for females, whereas BMI remains the most important predictor for males.

Furthermore, our findings highlight the importance of gender-specific model selection in diabetes risk prediction, with the Bayesian Additive Regression Tree (BART) proving to be the most effective model. Future research could explore the geographical variations in diabetes risk across different regions of India, considering the diverse dietary patterns. Additionally, integrating dietary habits, genetic predisposition, and socio-economic status into predictive models will allow for a more comprehensive assessment. Advanced Bayesian non-parametric models and deep learning techniques could further enhance accuracy and enable personalized risk stratification.

References

1. IDF Diabetes Atlas (2021), 10th edition. https://diabetesatlas.org/atlas/tenth-edition/.
2. Y. Khader, A. Batieha, H. Jaddou, M. El-Khateeb, K. Ajlouni, The performance of anthropometric measures to predict diabetes mellitus and hypertension among adults in Jordan, BMC Public Health. **19**, 1416 (2019). https://doi.org/10.1186/s12889-019-7801-2.
3. A. Awasthi, C. R. Rao, D. S. Hegde, N. K. Rao, Association between type 2 diabetes mellitus and anthropometric measurements - a case control study in South India, J Prev Med Hyg. **58 (1)**, E56 (2017).
4. H. E. Bays, R. H. Chapman, S. Grandy, the SHIELD Investigators' Group, The relationship of body mass index to diabetes mellitus, hypertension and dyslipidaemia: comparison of data from two national surveys, Int J Clin Pract. **61 (5)**, 737-747 (2007). https://doi.org/10.1111/j.1742-1241.2007.01336.x.
5. M. Ashwell, P. Gunn, S. Gibson, Waist-to-height ratio is a better screening tool than waist circumference and BMI for adult cardiometabolic risk factors: systematic review and meta-analysis, Obes Rev. **13**, 275-286 (2012). https://doi.org/10.1111/j.1467-789X.2011.00952.x.
6. M. Ashwell, S. Gibson, Waist-to-height ratio as an indicator of 'early health risk': simpler and more predictive than using a 'matrix' based on BMI and waist circumference, BMJ Open. **6 (3)**, e010159 (2016). https://doi.org/10.1136/bmjopen-2015-010159.
7. Sujata, R. Thakur, Unequal burden of equal risk factors of diabetes between different gender in India: a cross-sectional analysis, Sci Rep. **11**, 22653 (2021). https://doi.org/10.1038/s41598-021-02012-9.
8. C. W. Chia, J. M. Egan, L. Ferrucci, Age-Related Changes in Glucose Metabolism, Hyperglycemia, and Cardiovascular Risk, Circ Res. 123(7), 886-904 (2018). https://doi.org/10.1161/CIRCRESAHA.118.312806.
9. A. Gelman, A. Jakulin, M. G. Pittau, Y. Su, A weakly informative default prior distribution for logistic and other regression models, Ann. Appl. Stat. **2 (4)** 1360 – 1383 (2008). https://doi.org/10.1214/08-AOAS191.

10. H. A. Chipman, E. I. George, R. E. McCulloch, BART: Bayesian Additive Regression Trees, Ann. Appl. Stat. **4 (1)**, 266-298 (2010). https://doi.org/10.1214/09-AOAS285.
11. R. Jain, D. Bhattacharya, M. Das Gupta, S. Banerjee, Lifestyle, BMI, Age and Waist-to-Height Ratio as Indicators for Type 2 Diabetes Mellitus: A Gender Based Comparative Study in Kolkata, West Bengal, India, Asian Journal of Statistical Sciences, 3 (2), 169-176 (2023).