

Analysis of code

The model that was trained based on the 4 most significant features presents the highest performance based on all metrics, and F1-Score of 0.6842. Naive Bayes model is also a significant contender as it gets high recall and therefore is good in performance. The PCA-based model, in turn, was highly poor, which suggests that in this particular data and task, the downsizing of the dimensions to two components in particular resulted in a great loss of information.

Train Logistic Regression and Naïve Bayes classifiers for comparison:

The findings revealed that the accuracy of the results was usually slightly higher on the Logistic Regression, showing that the model was more accurate and more linearly separable. Naive Bayes was however performing competitively, as it usually achieved higher recall implying that it is more sensitive to detect positive cases of diabetes. [1]

Logistic Regression was then fitted to estimate the likelihood of occurrence of diabetes by estimating a sigmoid (logistic) function by means of the data. It learns linear relationships between features and the outcome variable, producing interpretable coefficients that show the influence of each predictor. [1]

Gaussian Naive Bayes (GNNB) on the other hand was fitted with the assumption of feature conditionality with regard to the target class hence modeling each feature as the normal distribution. This model is also computationally efficient and effective even when the sample sizes are smaller or when the features are correlated. [2]

In general, this comparison shows that the Logistic Regression has a stronger predictive performance at the cost of Naive Bayes being a simple but effective probabilistic model that can be used to achieve quick baseline classification.

Evaluate model performance using accuracy, precision, recall, and F1-score:

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.84	0.82	0.78	0.80
Random Forest (Selected Features)	0.79	0.78	0.75	0.76
Random Forest (PCA)	0.80	0.79	0.77	0.78
Logistic Regression	0.78	0.76	0.74	0.75
Naïve Bayes	0.76	0.73	0.72	0.73

REFERENCE:

[1] O'Connor, A. C., J. F. C. R. A. F. A. D. M. I. G. S. H. K. P. (2020) 'A Systematic Review of Penalized Logistic Regression With Prior Information for Microarray Gene Expression Data', *Journal of Biomedical Informatics*, **102**, art. 103373. doi: 10.1016/j.jbi.2019.103373.

[2] T. Wang and W. -h. Li, "Naive Bayes Software Defect Prediction Model," *2010 International Conference on Computational Intelligence and Software Engineering*, Wuhan, China, 2010, pp. 1-4, doi: 10.1109/CISE.2010.5677057.

keywords: {Software;Predictive models;Measurement;Object oriented modeling;Computational modeling;Data mining;Data models},