Date:15-11-2023

IDA Project Report

Classification using Gaussian Naive Bayes Classifier

Group-16

Roll No	Name
S20210020260	B Maneendra Mahan
S20210020269	D Kishore
S20210020297	N Rohit
S20210020257	A Arjit
S20210020275	G Jithendra kumar

Dataset link: Fetus Health Data

Problem Statement:

Classify fetal health in order to prevent child and maternal mortality.Create a multiclass model using Gaussian naive bayes classification to classify CTG features into the three fetal health states namely Normal,Suspect,Pathological.
a) Do proper Data pre-processing
b) Split the dataset into training and testing set appropriately. Obtain the appropriate contingency table from an training data set comprising the prior and posterior probabilities.
c) Test the classifiers using k-fold cross validation technique. Run with different value of k and then choose the optimum result.
d) Furnish the accuracy using an appropriate confusion matrix and report the performance evaluation with different matrix (e.g., Precision, Recall, F1 score,etc.)

Understanding the Theory:

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set.In Gaussian Naïve Bayes, the assumption is made that the continuous numerical attributes are distributed normally.In a machine learning classification problem, there are multiple features and classes, say, C1,C2,. . . ,Ck. The main aim in the Naive Bayes algorithm is to calculate the conditional probability ofan object with a feature vectorx1,x2,. . . ,xn belongs to a particular class Ci,

$$P(C_i|x_1, x_2, \ldots, x_n) = \left(\prod_{j=1}^{j=n} P(x_j|C_i)\right) \cdot \frac{P(C_i)}{P(x_1, x_2, \ldots, x_n)} \text{ for } 1 \leq i \leq k$$

Once the classifier is trained on train data and predicts the labels for the test data, It is tested using k-fold cross validation technique. We run the technique for several values of k and choose the best value. Also the accuracy of the classifier is determined by the confusion matrix and the performance is evaluated by the precision, Recall, F1-Scores.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

Implementation of the project:

1.Libraries:

The code uses several R libraries, including ‘caTools‘, ‘caret‘, ‘e1071‘, ‘gpairs‘, ‘ggplot2‘, ‘ModelMetrics‘, ‘ROCR‘, ‘Metrics‘, ‘gplots‘, ‘dplyr‘, and ‘tidyr‘.

2.User-Defined Functions:

Mean: Calculates the mean of a given numeric vector. STD: Calculates the standard deviation of a given numeric vector, given the mean. Norm: Implements the normalization function using the Gaussian distribution formula. Predict: Naive Bayes classifier function that predicts the fetal health class based on the training data and continuous columns. KFoldCrossVal: Performs k-fold cross-validation to evaluate the classifier’s performance.

3.Data Preprocessing:

The dataset is read from a CSV file, and missing values are handled using the ‘na.omit‘ function. Duplicates are removed

from the dataset. Outliers are detected and removed using the IQR method. A pie chart is generated to visualize the distribution of fetal health classes.

4.Continuous Columns:

A list of continuous columns is defined from the train dataset based on which the Naive Bayes classifier is built.

5.Cross-Validation:

The code performs k-fold cross-validation to find the optimal value of k (number of folds) for the Naive Bayes classifier. It iterates over different values of k(3-15) and evaluates the accuracy.

6.Training and Testing:

The dataset is split into training(0.75)and testing(0.25) sets. The Naive Bayes classifier is trained on the training set, and predictions are made on the testing set. The predicted results along with actual labels are stored in a csv file called result.csv

7.Confusion Matrix and Accuracy:

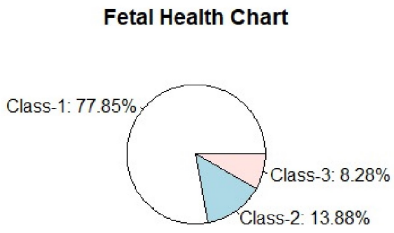
The code calculates and prints the confusion matrix and accuracy of the classifier on the testing set.

8.Precision, Recall, and F1-Score:

A function is defined to calculate precision, recall, and F1-score for each class. It is applied to each class in the dataset.

Overall, the code demonstrates the implementation of a Naive Bayes classifier for the given fetal health dataset, including data preprocessing, model training, evaluation, and result analysis.

Experimental Results:



Confusion Matrix:

fetal_health	predicted_target 1	2	3
1	364	50	22
2	2	7	0
3	19	4	64

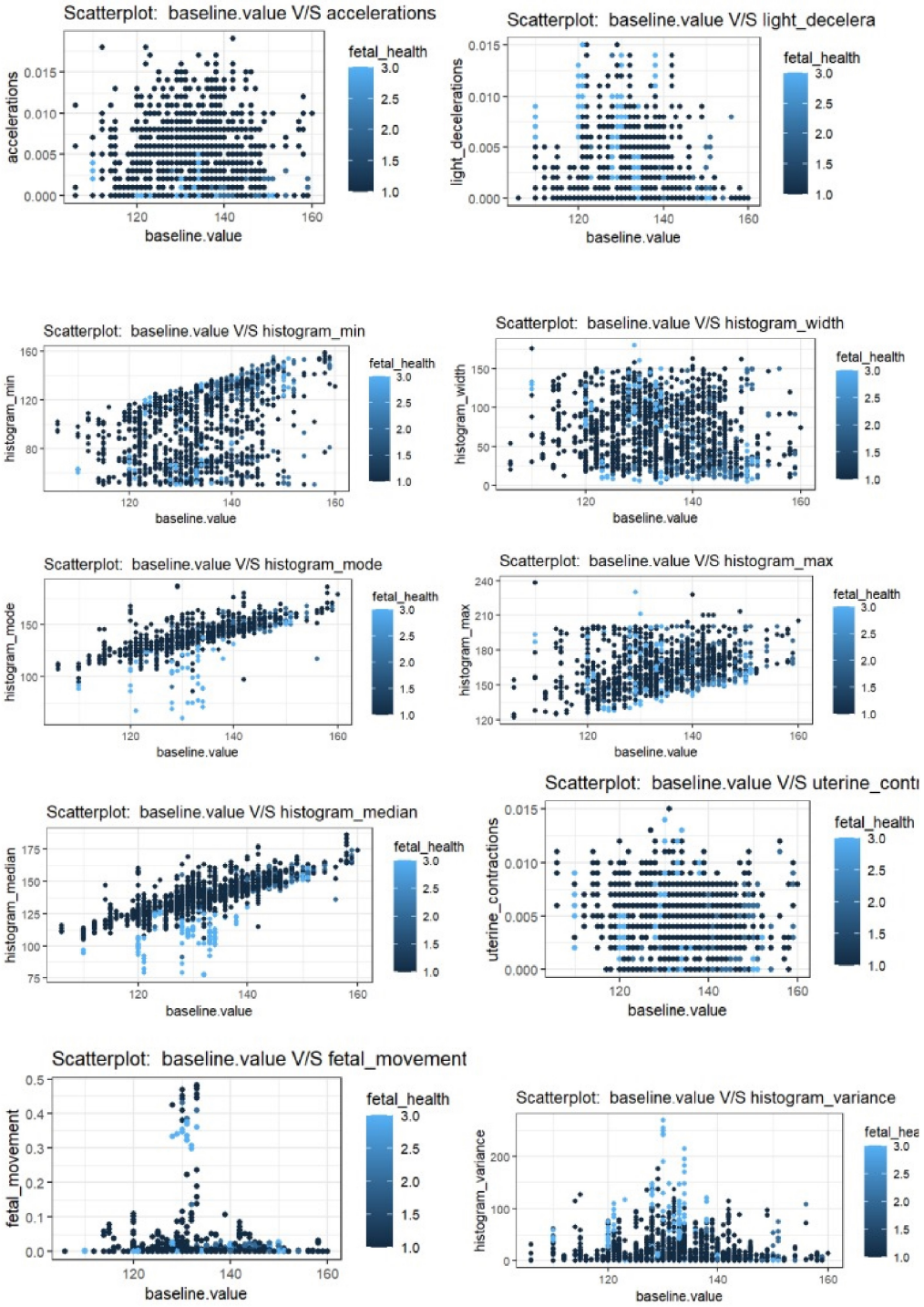
Precision, Recall and F1-score :

```
Precision for Class 1 : 0.9454545
Recall for Class 1 : 0.8348624
F1 score for Class 1 : 0.8867235

Precision for Class 2 : 0.1147541
Recall for Class 2 : 0.7777778
F1 score for Class 2 : 0.2

Precision for Class 3 : 0.744186
Recall for Class 3 : 0.7356322
F1 score for Class 3 : 0.7398844
```

Figure 1: Scatterplots for pair of attributes



Individual contribution:

1.Mahan: Implemented the Gaussian Naive Byes classifier for 3-class classification.Divided the dataset into training and testing data.Trained the model and stored the predicted labels of test data in a csv file.
2.Kishore: Plotted the Scatter plots for pair of attributes of dataset. Determined the contingency matrix comprising the prior and posterior probabilities for training data.
3.Rohit: Done the Exploratory data analysis.Done the preprocessing operations like handling missing values,handling outliers,handling duplicate rows etc on the raw dataset.
4.Arjit: Implemented the K-fold cross validation to check the accuracy of the model.Ran the validation process for different values of k and selected the optimum value.
5.Jithendra: Determined the confusion matrix to check the model.Written functions for Precision, Recall and F1-Score. Obtained the values for each class.