Predicting Mortality Rate in ICUs due to Heart Attack

Name:	**Anirudh Tikoo**
Registration No./Roll No.:	20043
Institute/University Name:	IISER Bhopal
Program/Stream:	DSE
Problem Release date:	August 17, 2023
Date of Submission:	November 19,2023

1 Introduction

The data has 48 features and 1058 instances taken from different age groups, weight, medical test results having 2 classes 0 for Non Fatal and 1 for Fatal. .There are 915 cases of data belong to non fatal class label and 143 belong to fatal class label.There are 1723 empty data points and I imputed the median of the feature for the empty values.The given problem is a **Classification Problem** as the class labels are countable and discrete.

2 Methods

First, the CSV files was uploaded to Jupyter Notebook and converted it into a data frame. The data had null values which were filled by calculating the median of each feature and replacing the null values with the median.There was categorical data in the dataframe which was encoded into a binary vector using OneHotEncoder.Finally we replaced the categorical values with their binary vectors and used SelectKBest to select the best 34 features from the non categorical data using Mutual Information as the selection parameter.It selects the features based on the mutual dependencies between the target variables.Finally Principal Component Analysis was applied on the whole dataset and 29 features were selected which gave minutely better results than just applying SelectKbest.The data is also transformed and normalized.The data was split into training and test set using StratifiedKFold into 75percent training set and 25percent test set.The classifiers that were used were MultilayerPerceptron which is a type of neural network for non linearly separable data,K Nearest Neighbor, Logistic Regression, Random Forest, Support Vector Machine and Gaussian Naive Bayes.All it's parameters were found using GridSearchCV which exhaustively runs over all the parameters described in the parameter grid. Finally the fit was tested for all classifiers and the result was evaluated. The github repository is https://github.com/Anmitee/project1.git

3 Experimental Setup

The training set was tested on different classifiers and their best parameters were found using GridSearchCV.

K Nearest Neighbours was fine tuned on parameters like neighbors to define the value of k.It also has weights parameter on how the weight is to be calculated and it was taken to be uniform i.e. uniform weight was assigned to each to all data points.Algorithm used to compute the nearest neighbor .Leaf size is passed as 5,his can affect the speed of the construction and query, as well as the memory required to store the tree.P is used to define the metric in which the distance is measured and it was taken as 3.0 here which means Minkowski distance is taken. Minkowski distance or is a metric in a normed vector space which can be considered as a generalization of both the Euclidean distance and the Manhattan distance.

SVC is the Support Vector Machine for Classification.The parameters for SVC are Regularization Parameter or C controls the trade off between maximizing the margin and minimizing misclassification.It is used with soft margins for linearly inseperable data.Kernel decides the type of kernel to be used like RBF which is chosen by default.Here kernel is taken as sigmoid.Gamma decides the kernel coefficient for the function chosen by kernel.The independent term in kernel is 0 and decision function shape is one vs one.

Gaussian NB has parameter variable smoothing which portions of the largest variance of all features that is added to variances for calculation stability here the value is 1e-07.

Random Forest Classifier The variable n estimators is used to determine the number of trees in the forest.The parameter criterion measures the quality of the split in a tree which has been taken as entropy. Max depth define the maximum depth of a tree and The number of features to consider when looking for the best split is defined by Max features.CCP Alpha is the complexity parameter used for Minimal Cost-Complexity Pruning. The subtree with the largest cost complexity that is smaller than ccp alpha will be chosen.

Logistic RegressionIn Logistic regression penalty specifies the norm of the penalty. C specifies Inverse of regularization strength; must be a positive float. Like in support vector machines, smaller values specify stronger regularization.Solver is the algorithm to use in the optimization problem.Newton-cg is used here for the solver.The evaluation scores are discussed in the next part.

4 Results and Discussion

The tables created to show the experimental results.

Table 1: Performance Of Different Classifiers Using All Features

Classifier	Precision	Recall	F-measure
Adaptive Boosting(Logistic Regression)	0.549	0.616	0.55
Multilayer Perceptron	0.54	0.56	0.55
K-Nearest Neighbor	0.500	0.504	0.507
Logistic Regression	0.564	0.649	0.577
Random Forest	0.51	0.52	0.63
Support Vector Machine	0.463	0.462	0.465
Gaussian NB	0.58	0.57	0.625

Table 2: Confusion Matrices of Different Classifiers

Actual	Predicted Class	
Class	0	1
0	216	29
1	12	7

Adaptive Boosting

Actual	Predicted Class	
Class	0	1
0	211	33
1	17	3

K-Nearest Neighbor

Actual	Predicted Class	
Class	0	1
0	211	31
1	17	5

Multipayer Perceptron

Actual	Predicted Class	
Class	0	1
0	157	14
1	71	22

Logistic Regression

Actual	Predicted Class	
Class	0	1
0	225	34
1	3	2

Random Forest

Actual	Predicted Class	
Class	0	1
0	192	33
1	36	3

SVM

Actual	Predicted Class	
Class	0	1
0	216	29
1	12	7

Gaussian NB

5 Conclusion

The results found were not optimal as in many cases the Fatal class i.e. class 1 got misclassified or sometimes did not even get classified.This decreased the overall f1-score.Mutual Information quantifies the information gained about one variable by observing the other one.This is helpful in this dataset as instances with class 1 are very less hence we are using class 0 as well to determine which instances lie in class 1.Also f1-macro is used as a method of evaluation as f1-micro is giving extremely high false values even for cases where class 1 isn't even classified in the test.Therefore the metrics used to evaluate the models are very important