UE17CS303 – MACHINE LEARNING

ASSIGNMENT

PES1201701629 PRUTHVISH E

PES1201701435 HARISH P B

PES1201700649 ASHRITH SRIDHAR

Low Birth Weight Classification

Classifying newborns as having LBW or Not-LBW

1. Problem Statement:

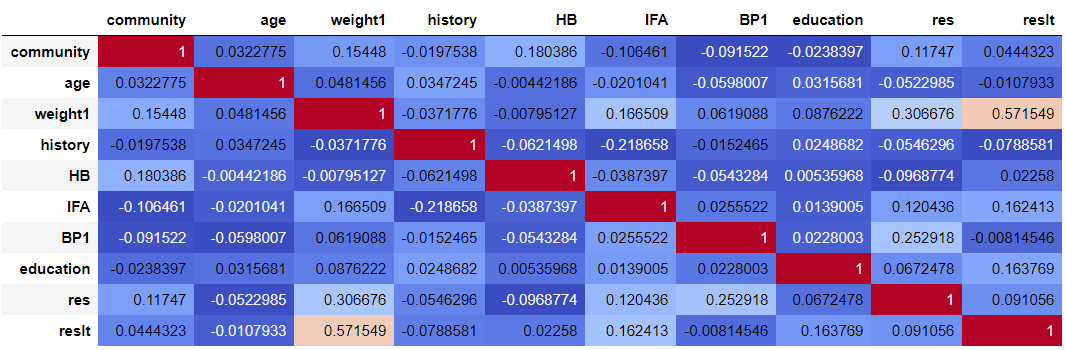
The objective of this endeavor is to apply ML techniques on the given dataset of LBWs, whose data has been taken from Kurnool district of Andhra Pradesh

* The dataset contains 10 columns which are community, age, weight1, history, HB, IFA, BP1, education, res, reslt.

Data Preprocessing

* dropped education column
* statistical imputation of missing data with mean, median or mode
* divide the dataset into two dataframes based on label and imputation on each dataframe is performed
* Replace NAN values with median in case of skewed data
* Replace NAN values with mean otherwise
* Replace NAN values with mode in case of categorical attributes

correlation matrix for the cleaned dataset



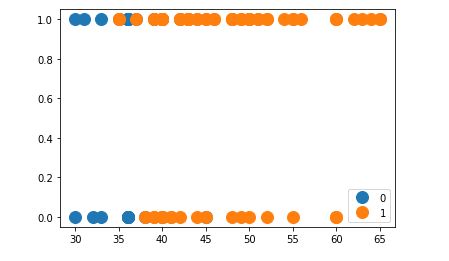
1. ML Techniques employed

Weighted KNNs

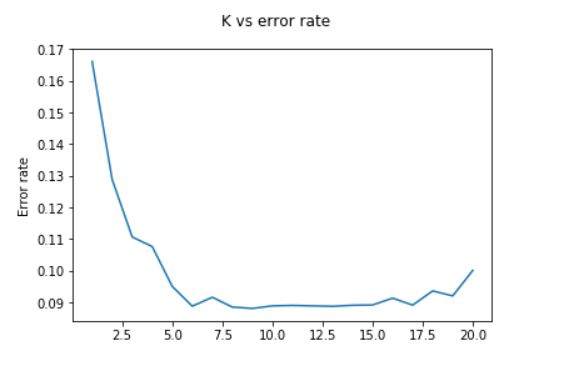
The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other.

Weighted kNN is a modified version of [k nearest neighbors](https://www.geeksforgeeks.org/k-nearest-neighbours/). The weighted k-nearest neighbors (K-NN) classification algorithm is a relatively simple technique to predict the class of an item based on two or more numeric predictor variables. At the end, by voting, the prediction is deduced.

The weights are assigned based on the correlation matrix.

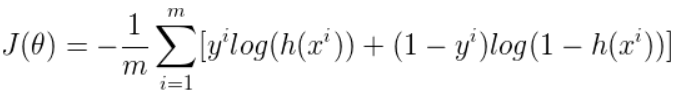


**Choosing the right K**



Logistic Regression

* Logistic Regression is a classification algorithm that is used to predict the probability of a categorical dependent variable
* In logistic regression, the dependent variable is a binary variable that contains data coded as 1 or 0
* the activation function used is in our case is sigmoid function
* the objective is to minimize the cost function with gradient descent



Bagging

Bagging is an abbreviation for "bootstrap aggregating". It’s meta-algorithm, which takes M subsamples (with replacement) from the initial dataset and trains the predictive model on those subsamples. The final model is obtained by averaging the "bootstrapped" models and usually yields better results.

Bagging is a technique to take models with high variance and low bias and reduce their variance without decreasing their bias.

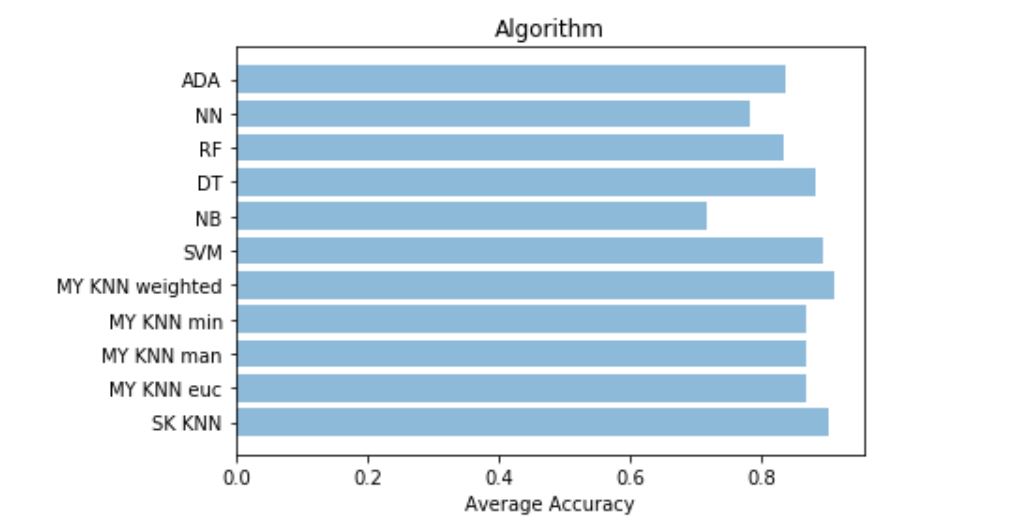
1. **Summary of results**

Average Accuracy of Weighted KNN’s using 5 fold validation on multiple trials was 92%.

Average Accuracy of Logistic Regression using 5 fold validation on multiple trails was 90%

Average Accuracy of Bagging using 5 fold validation on multiple trails was 92%

Comparison with different ML models of sklearn :-



1. **Analysis and Conclusion:**

Weighted KNN’s perform the best for classifying the data whether its LBW or not LBW as related labels stay close together in the 9th dimensional space.

We can see that weights are clearly skewed as weight1 feature contributes the most of all the features.

Hence neighbors with similar weight1 value will belong to the same target values, the only case where it is not valid is when weights of all other columns sum to be greater than the weight1 value.