IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

COMPARING CLASSIFICATION MODELS FOR MULTILABEL IMBALANCED FETAL GROWTH DATA

Preethi Jayarama Shetty

Corresponding author: Department of statistics, Mangalore University, Karnataka

Abstract: In humans, an unborn baby that develops and grows inside the uterus (womb). The fetal period begins eight weeks after fertilization of an egg by sperm and ends at the time of birth. Staying healthy at the time of pregnancy is most important for fetal growth, otherwise baby's body and organs don't grow as much as they should. Growth of the fetal is affected by maternal factors such as maternal size, weight, nutritional state, anemia, cigarette smoking, alcohol consumption and many more. Also, placental factors such as size, microstructures, umbilical blood flow, transporters and binding proteins, nutrient utilization and nutrient production and fetal factors like fetus genome, nutrient product and hormone output are affecting. Handling of imbalanced data processes is great challenges for the researchers to identify the class labels. In health science, accurate detection of class labels using classification model will reduce burden of doctors as well as avoid great loss. The main focus of the study is to identify the major factors associated with labels of fetal growth and identifying the better classification model to handle multinomial imbalanced fetal growth data. The prediction performance of different classification models under SMOTE algorithm is compared based on evaluation measures and better classification model to deal with imbalanced data is reported.

Keywords: Imbalanced data, multinomial data, SMOTE, Prediction.

1. Introduction:

Pregnancy is a period during which offspring develop inside a woman . Fetal health is the indication of proper growth of the fetus in the pregnant woman's uterus during the gestation period. In humans, an unborn baby that develops and grows inside the uterus (womb). The fetal period begins eight weeks after fertilization of an egg by sperm and ends at the time of birth. Staying healthy at the time of pregnancy is most important for fetal growth, otherwise baby's body and organs don't grow as much as they should. Growth of the fetal is affected by maternal factors such as maternal size, weight, nutritional state, anemia, cigarette smoking, alcohol consumption and many more. Also, placental factors such as size, microstructures, umbilical blood flow, transporters and binding proteins, nutrient utilization and nutrient production and fetal factors like fetus genome, nutrient product and hormone output are affecting. Although decreasing of fetal movement due to internal and external factors is the crucial causes that adverse birth outcome unless monitored consistently. In order to examine the fetal well being UCI CTG dataset is the standard source of information and an electronic fetal state monitoring that used for predictive classification purpose. The FHR has been obtained from a Doppler ultrasound and mother's UC by pressure transducer were recorded on cardiotocograph (CTG).

Mothers related high blood pressure abnormality during the gestation period suffer 10% of pregnant women around the world that leads to stress, impairment, and death of both lives as the World Health Organization (WHO) studies reflect. This affects the babies in uterus to get insufficient blood circulation, which reduces fetus movement. Data mining and machine learning algorithms were used for maximizing performance of choosing classifier that builds an accurate learning model for predicting the risk based on CTG dataset. The baby's Fetal Heart Rate and UC are collected on CTG techniques that is highly instrumental in the previous abnormalities identification and provides the obstetrician to predict future risks.

2. Literature review:

C. A. Prajith, et.al (2016), described the growth of scar tissue due to inflammation, infection, or injury so called liver fibrosis. This disease could be the reason for liver cirrhosis. The use of various non-invasive imaging techniques was quite common for the treatment of liver fibrosis. These techniques included MRI, CT, Electrography, and ultrasound. This study was focused on the extraction of texture features from liver images of ultrasound. This work implemented various classification models such as ANN, GMM, and SVM for classifying the risk level of the liver fibrosis. SVM has a specificity of 95 % , the sensitivity of 93.33%, and an accuracy of 94 % .

Thirunavukkarasu, et.al(2019), studied, the medical field produced a huge amount of healthcare data on a daily basis. The use of the ML algorithm was quite common for finding concealed information for disease detection and efficient decision-making. Significant growth in Liver diseases had been noticed with the time. In several countries, these diseases were the major cause of death. This work was focused on the prediction of liver disorder with the help of several classification algorithms. These algorithms included LR, KNN, and SVM. In this work, the comparison of these algorithms had been carried in terms of accuracy rate and confusion matrix.

Khair Ahammed et.al. (2020). In this work, a machine learning based model has been proposed that can classify hepatitis C virus infected patient's stages of liver. Researchers gathered the instances of liver fibrosis disease of Egyptian patients from UCI machine learning repository. To balance instances of multiple categories, synthetic minority oversampling methodology was used. Later, different feature selection methods to identify significant features of hepatitis C virus in this dataset. KNN was used to classify hepatitis C virus infected patients. This result has been useful to scrutinize and take decision in hepatitis C virus infectious disease. In this experiment, HCV patient's records was investigated using machine learning techniques to detect the stages of the HCV patient. Top five best sorted results of different classifiers based on various evaluation criteria. In this circumstance, GB, NB and LR shows their accuracy less than 50%, so they are discarded in this section. Thus, SMOTE was applied 8 times in the raw HCV dataset and generated more synthetic instances like existing instances. Then, various feature selection techniques were used to the SMOTE generated HCV patient's instances and produce datasets.

Golmei Shaheamlung, Harshpreet Kaur (2021), the Diagnosis of Chronic Liver Disease using Machine Learning Techniques. In this research work is based on liver disease prediction using machine learning algorithms. Liver disease prediction has various levels of steps involved, pre-processing, feature extraction, and classification. In this s research work, a hybrid classification method is proposed for liver disease prediction. And Datasets are collected from the Kaggle database of Indian liver patient records. The proposed model achieved an accuracy of 77.58%. The proposed technique is implemented in Python with the Spyder tool and results are analyzed in terms of accuracy, precision, and recall

3. Materials and Methods

In real life, most of the categorical dataset are imbalance. Handling of imbalanced data possesses great challenges for the researchers to identify the class labels. In health science, correct detection of presence of a disease using good classification model will reduce burden of doctors as well as avoid great loss. In this study, we handle imbalanced liver disease data using SMOTE and ROSS sampling techniques.

3.1 ROSS: Random over-sampling is oversampling technique used to balance the imbalanced dataset. In ROSS, new minority samples are created by randomly selecting training samples from minority class, and then duplicating it. In doing therefore, the category distribution is often balanced, however this could typically cause over-fitting and longer training time throughout imbalance learning method.

3.2 SMOTE: To overcome the issue of over-fitting and extend the decision area of the minority class samples, a novel technique SMOTE ''Synthetic Minority Oversampling Technique'' was introduced by Chawla, this technique produces artificial samples by using the feature space rather than data space. It is used for oversampling of minority class by creating the artificial data instead of using replacement or randomized sampling techniques. It was the first technique which introduced new samples in the learning dataset to enhance the data space and counter the scarcity in the distribution of samples. The oversampling technique is a standard procedure in the classification of imbalance data (e.g., minority class).

4. CLASSIFICATION MODES:

4.1 DECISION TREE:

A **decision tree** could be a flowchart-like tree structure, where the uppermost node in tree is the root node, test on an attribute denotes by every **internal node**, an outcome of the test represented by every **branch**, and each terminal node holds a class label. Given a tuple which associated unknown class label, the attribute values of the tuple are tested against the choice tree. A path is traced from the root to a terminal node, that holds the class predicted values for that tuple. This is appropriate for exploratory knowledge discovery since construction of decision tree doesn't need any domain knowledge or parameter setting. Decision trees can handle multidimensional data. The average of dependent variable values in a tuple is taken as the predicted value for all those tuples. CART algorithm applies the Gini index as the attribute selection measure.

4.2 RANDOM FOREST

Random forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. To correct overfitting problem of decision tree, Random decision forest is the alternative. Random forest adds additional randomness to the model, while growing the trees. Random forest searches for the best feature among a random subset of features instead of searching for the most important feature while splitting a node. Therefore, in random forest, only a random subset of the features is taken into consideration by the algorithm for splitting a node. Another great quality of the random forest algorithm is that it is very easy to measure the importance of each feature on the prediction. By looking at the feature importance you can decide which features to possibly drop because they don't contribute enough to the prediction process. The hyperparameter in random forest are used to increase the predictive power of the model as well as the speed of model.

4.3 SUPPORT VECTOR MACHINE:

SVM is a supervised machine applicable for classification and regression problems. This algorithm creates a hyperplane which separates the data into two classes. SVM is an algorithm that takes the data as an input and outputs, a line that separates those classes if possible. According to the SVM algorithm, the data points which are closest to the line from both the classes are called support vectors. The aim is to maximize the margin, that is distance between the line and the support vectors. The maximize margin is nothing but optimal hyperplane. Thus, SVM tries to create a choice boundary in such the way that the separation between the 2 categories is as wide as possible.

4.4 NAÏVE BAYES CLASSIFIER:

Bayesian classifiers are statistical classifiers. This is the algorithm to predict the probability of given tuple that belongs to a particular class. Bayesian classification is based on Bayes' theorem. A simple Bayesian classifier is known as the 'naïve Bayesian classifier' to be comparable in performance with decision trees and selected neural network classifiers. When naïve Bayesian classifier applied to a large database, it gives high accuracy and speed. According to this algorithm, effect of an predictor on a given class is purely independent of the values of the other attributes. It is known as class conditional independence. To predict the class label of X, P(X|Ci)P(Ci) is evaluated for each class Ci. The classifier predicts that the class label of X is Ci if and only if it is the class that maximizes P(X|Ci)P(Ci).

4.5 K-NN CLASSIFICATION:

KNN compare the given test tuple with training tuple which is similar to it, so we can say it is based on learning by analogy. Once given an unknown tuple, a K- nearest neighbour classifier search the pattern area for the k-training tuple that area unit nearest to the unknown tuple. Closeness is outlined in terms of distance matric. For K-nearest neighbour classification, the unknown tuple is allotted to the foremost category among its K-nearest neighbours. K-nearest neighbour classifiers can also be used for prediction, that is, to return a real valued prediction for a given unknown tuple. A good value for K, number of nearest neighbours, can be found experimentally or k may be taken as

$k = \sqrt{number\ of\ training\ tuples}$

4.6 NEURAL NETWORK

The method of neural networks training is based on some initial parameter setting, weight, bias, and learning rate of algorithm. It starts its leaning with some initial value and weight gets updated on each iteration. The training of neural network is time consuming and it structure is complex. These feature made neural network less suitable for classification in data mining. Some method can be proposed to learn both the network structure and updating the weight. Adjustment of weight in ANN is combinatorial problem and to find the desired output we have to optimize the weight. Some learning methods for ANN in different classification problem are as follows:

a) Artificial neural network with back propagation

One variant of ANN with BP is proposed in give application of neural network for classification of Landsat data. The back propagation algorithm is used for training of neural network. Other variant of ANN with BP is proposed in is used for multispectral image classification. The BP is trained on classical area of image and then the neural network is used to classify the image.

b) Improved back propagation algorithm Discuss the training of neural network with back propagation algorithm using gradient delta rule. It is highly applicable for parallel hardware architecture. Rather than being held constant the momentum factor is determined on each step. Compared to conventional BP improved back propagation has better speed and convergence stability.

Soft computing contains some meta-heuristic algorithms like cuckoo search, firefly algorithm, genetic algorithm, particle swarm optimization. These meta-heuristic algorithms can be used for training of neural network. The meta-heuristic algorithms produce approximate result and applicable to any field. These algorithm are used where traditional algorithm produce local optimum. Traditional algorithm also increase computational cost and use more time to produce result.

4.7 Classification table:

The confusion matrix is a method to evaluate accuracy of the logistic regression.

Predicted class / Actual class	Class 1	Class 2	Class 3
Class 1	True Negative (TN)	False Positive (FP)	True Negative (TN)
Class 2	False Negative (FN)	True Positive (TP)	False Negative (FN)
Class 3	True Negative (TN)	False Positive (FP)	True Negative (TN)

EVALUATION MEASURES:

- **Accuracy:** It is a measure that calculates the classifier's overall accuracy.
  It is formulated as:

$$\boldsymbol{Accuracy} = \frac{TN + TP}{\mathrm{TN} + \mathrm{FN} + \mathrm{FP} + \mathrm{TN}}$$

ANALYSIS AND DISCUSSION:

- Handling logit class data when the data is imbalance is difficult but when data has multiclass dependent variable along with imbalance is most challenging work for researchers. In this project we made a attempt to handle imbalanced multiclass data. The data has been collected from the website **https://www.kaggle.com/andrewmvd/fetal-health-classification** and it consists of 2126 instances with 12 chosen attributes which are multivariate datatypes. The variables under the study are as follows,
- baseline value:- FHR baseline (beats per minute)
- accelerations:- Number of accelerations per second
- fetal_movement:- Number of fetal movements per second
- uterine_contractions:- Number of uterine contractions per second
- light_decelerations:- Number of light decelerations per second
- severe_decelerations:- Number of severe decelerations per second
- prolongued_decelerations:- Number of prolonged decelerations per second
- abnormal_short_term_variability:- Percentage of time with abnormal short term variability
- mean_value_of_short_term_variability:- Mean value of short term variability
- percentage_of_time_with_abnormal_long_term_variability:- Percentage of time with abnormal long term variability
- mean_value_of_long_term_variability:- Mean value of long term variability
- fetal_health:-Tagged as 1 (Normal), 2 (Suspect) and 3 (Pathological)

All the independent variables are considered as significant variables since there is direct and indirect effect for the target variable.

Figure1: Correlation matrix

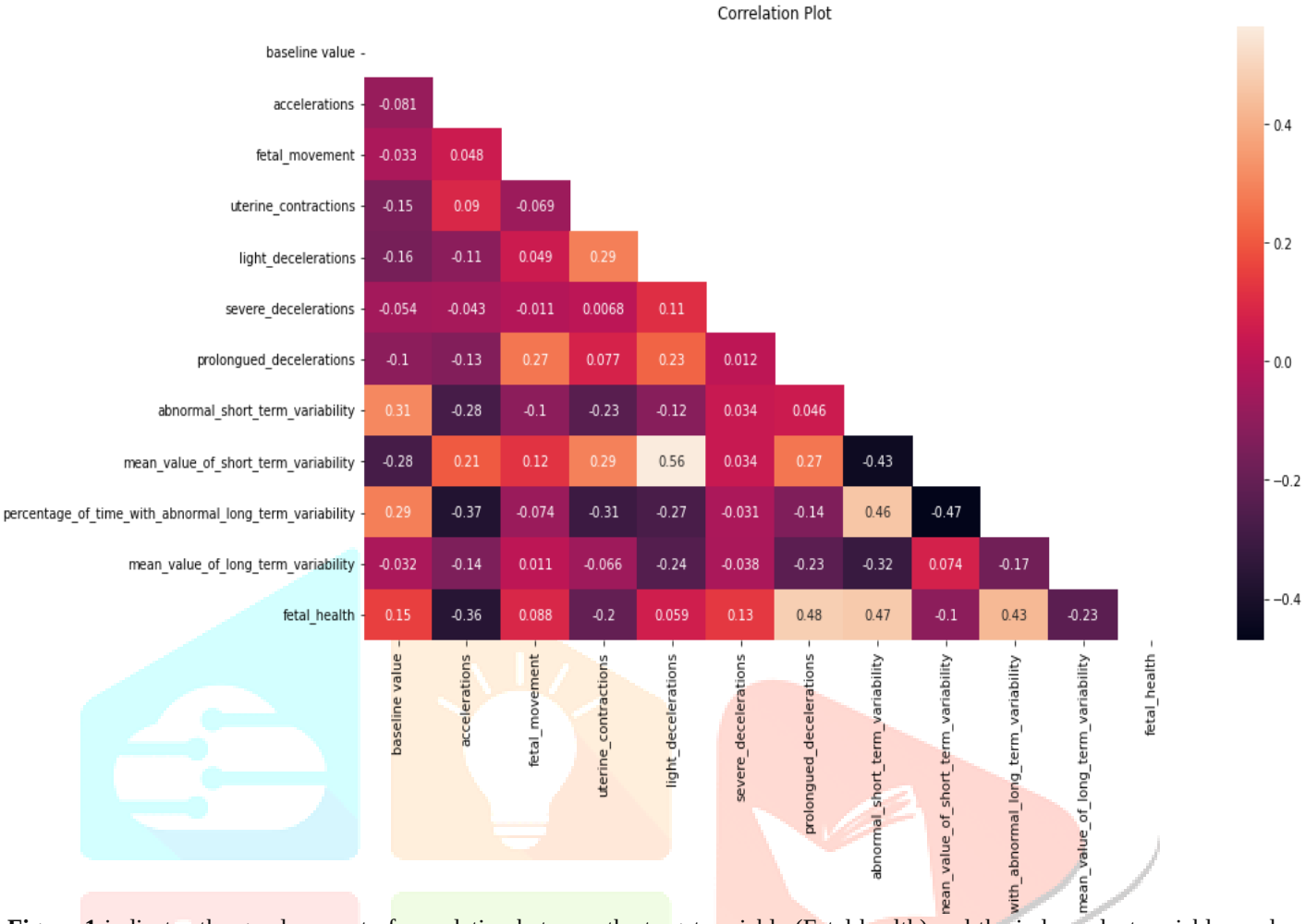
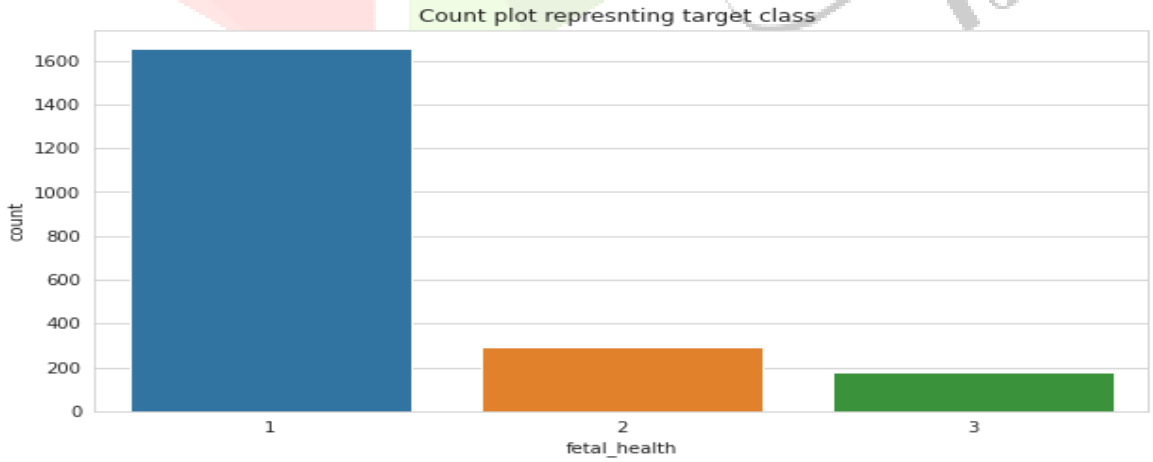


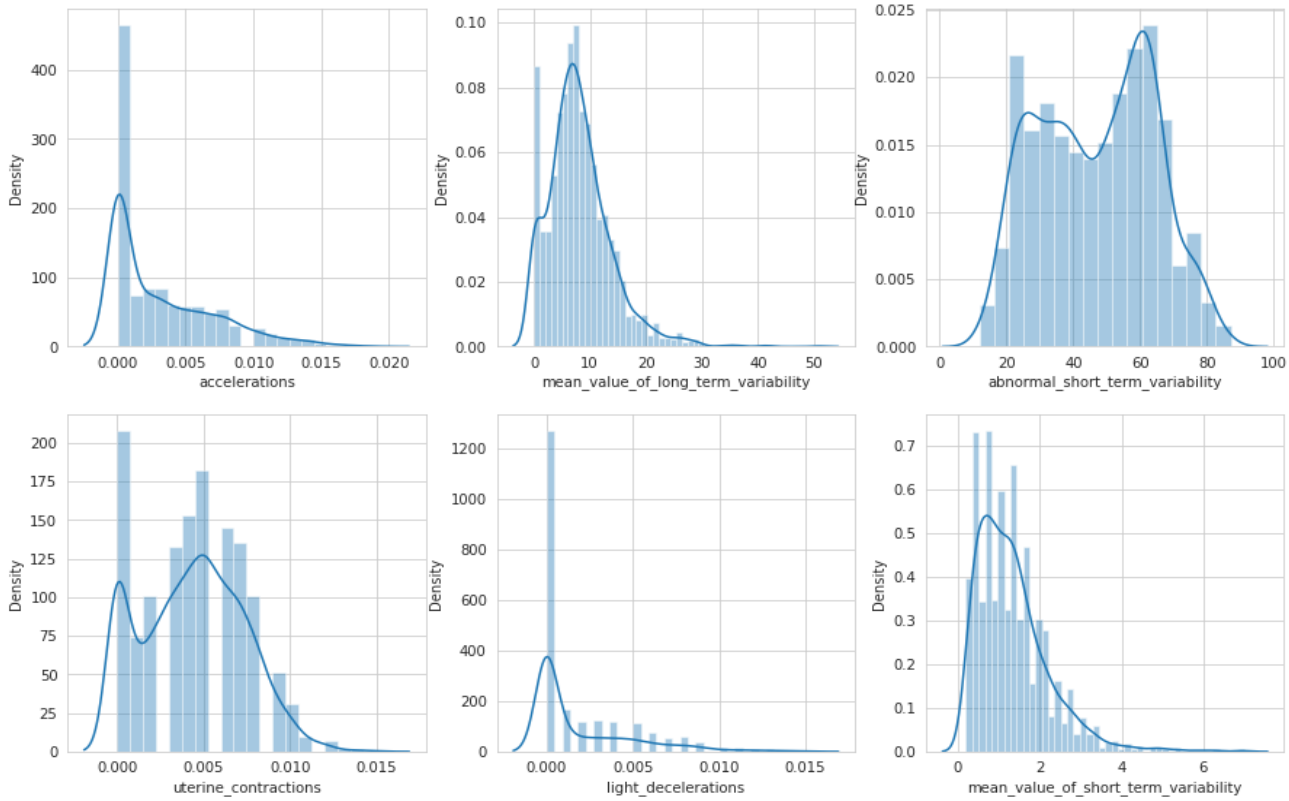
Figure 1 indicates the good amount of correlation between the target variable (Fetal health) and the independent variables and also, there is no multicollinearity between the independent variables.

Figure 2: Bar plot of Fetal health



From the figure 2, we observe that patients in Normal category is around 82% and the suspect & Pathological category are of 12% and 8% respectively. So, this is the indication of class imbalance

Figure 3: Distribution plot for continuous variables



- We observe that the acceleration variable is quiet right skewed, this may be due to the patients with the high value of acceleration low frequency.
- Light deceleration is quiet right skewed and we see most of the people miss the value zero so that high peaked.

Table 1: Performance evaluation of different classification models under different sampling techniques

Methods	Balanced accuracy		
	Class 1	Class 2	Class 3
RANDOM SAMPLING			
Decision Tree	0.8751	0.8459	0.89
Random Forest	0.9131	0.8814	0.96
KNN	0.8968	0.8648	0.8966
Neural Network	0.8793	0.8146	0.89405
Naïve Bayes	0.8682	0.821	0.78724
SVM	0.8909	0.8909	0.8909
SMOTE SAMPLING			
Decision Tree	**0.9825**	**0.9796**	**0.9798**
Random Forest	0.9379	0.837	0.9078
KNN	0.848	0.8065	0.805
Neural Network	0.8557	0.8818	0.8949
Naïve Bayes	0.9149	0.8719	0.8431
SVM	0.9232	0.9232	0.9232

From table 1, Decision tree under SMOTE sampling is better model for fetal health data.

Conclusion:

Handling of imbalanced data processes is great challenges for the researchers to identify the class labels. In health science, accurate detection of class labels using classification model will reduce burden of doctors as well as avoid great loss. The main focus of the study is to identify the major factors associated with labels of fetal growth and identifying the better classification model to handle multinomial imbalanced fetal growth data. The prediction performance of different classification models by considering imbalanced data and balanced data using SMOTE algorithm is compared based on evaluation measures. The result shows that prediction performance of all the six classifiers has been improved by considering SMOTE sampling techniques. Based on overall performance, Decision Tree under SMOTE over sampling is the best classifier to deal with imbalanced fetal growth data.

Reference:

- C. A. Prajith, A. Sukesh Kumar, Harish Kareem (2016) Supervised classification and prediction of fibrosis seriousness using ultrasonic images by IEEE Publisher.
- Ganesan, S. and Thirunavukkarasu, N (Dr.) (2019), Analysis of Importance and Purpose of Using Electronic Resources on Selected Engineering Colleges in Coimbatore District by Reshaping of Librarianship, Innovations and Transformation.
- Khair Ahammed, Md. Shahriare Satu, Md. Imran Khan, Md Whaiduzzaman (2020) Predicting Infectious State of Hepatitis C Virus Affected Patient's Applying Machine Learning Methods by IEEE Region 10 Symposium (TENSYMP), 5-7 June 2020, Dhaka, Bangladesh.
- Golmei Shaheamlung, Harshpreet Kaur (2021), the Diagnosis of Chronic Liver Disease using Machine Learning Techniques by IT in Industry, vol. 9, no.2, 2021.