Candidate Number: 28

PROJECT REPORT: ACIT4510 STATISTICAL LEARNING

INTRODUCTION

Cardiotocography also known as Electronic Fetal Monitoring EFM is a worldwide technique for monitoring the fetus. Fetal health is an important indicator of the well healthy being and overall development in the womb of the pregnant women during pregnancy. [1][19]

To reduce the child mortality is one of the several goals of the United Nation's Sustainable Development Goals to be achieved. This goal is one of the important factors of human progress. By 2030, United Nations expects that the countries should end the preventable deaths of the newborns as well as children under the age of 5. The goal is to decrease the under-5 mortality to as low as 25 per 1000 live births.[1][20]

Along with the child mortality is also maternal mortality which is 295000 deaths following pregnancy and childbirth as of 2017. Majority of these deaths nearly 94% had occurred in low resources conditions and could have been possibly prevented. [1][16]

Cardiotocograms (CTGs) are simple and cost effective option for monitoring and assessing fetal health. This helps the healthcare professionals to take further actions for preventing child and maternal mortality. Cardiotocography provides crucial information in knowing the fetuses that are suffering from lack of oxygen (i.e. Hypoxia). This is medically termed as fetal distress and needs medical intervention to prevent fetus death or other neurological diseases. [1] [16]

The equipment operates by sending the ultrasound pulses and then reading the response with the observation of fetal heart rate (FHR), fetal movements, uterine contractions and so on. [1]

The Fetal Health dataset is obtained with several features extracted from 2126 different CTG examinations. These were then classified by expert obstetricians into 3 different classes:[1]

- Normal
- Suspect
- Pathological

The aim of the project is to perform exploratory data analysis on the fetal health dataset and to predict the fetal state based on the input data from cardiotocography examination. Different machine learning classifier algorithms for classification problem would be experimented and implemented and each model would be evaluated. The goal is to build a predictive model to classify the CTG features into one of the three fetal health states that is normal, suspect or pathological with highest accuracy and least prediction error.

METHODS AND DATA

The dataset was obtained from the University of California Irvine Machine Learning Repository and from Kaggle dataset repository[1][2]. The data was collected from 2126

pregnant women who were in the third trimester of pregnancy and consisted of 22 attributes used in the measurements of FHR and uterine contractions (UCs) on CTG

An exploratory data analysis is done on the dataset using Jupyter Python for observing and finding interesting facts about the dataset.

Data Analysis and Visualization of Dataset

The fetal health dataset contains 2126 rows and 22 columns that is it contains 2126 records and 22 different features. The below table describes the features and their explanation.[16]

- *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
- *Prolonged 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
- *Histogram width*- width of FHR histogram
- *Histogram Min*- Low frequency of FHR histogram
- *Histogram Max*- High frequency of FHR histogram
- *Histogram Number of Peaks*- Number of peaks in FHR histogram
- *Histogram number of zeroes*- Number of zeroes in FHR histogram
- *Histogram mode*- Mode value in FHR histogram
- *Histogram mean*- Mean value in FHR histogram
- *Histogram median*- Median value in FHR histogram
- *Histogram variance*- Variance in FHR histogram- Histogram tendency- Histogram tendency : -1=left asymmetric;
  0=symmetric; 1=right asymmetric
- *Fetal Health*- Fetal Health classified as Normal = 1, Suspected = 2, Pathological = 3

Baseline value is estimated by an algorithm using fetal heart rate (FHR) and short term variability (STV) described in (Ayres-de-Campos et al. 2000). Acceleration is defined as increase in the FHR above the baseline for 15 to 120 seconds and deceleration is defined as the decrease in the FHR below the baseline for 15 to 120 seconds. Uterine contraction signal after filtering is evaluated for the contraction episodes. [2][21]

The table below presents the statistical analysis of the dataset for the numerical features

	Count	Mean	Std	Min	25%	50%	75%	Max
Baseline Value	2126.0	133.303857	9.840844	106.0	126.0	133.00	140.0	160.0
Accelerations	2126.0	0.003178	0.003866	0.0	0.000	0.002	0.006	0.019
Fetal Movement	2126.0	0.009481	0.046666	0.0	0.000	0.000	0.003	0.481
Uterine Contractions	2126.0	0.004366	0.002946	0.0	0.002	0.004	0.007	0.015
Light decelerations	2126.0	0.001889	0.002960	0.0	0.000	0.000	0.003	0.015
Severe decelerations	2126.0	0.000003	0.000057	0.0	0.000	0.000	0.00	0.001
Prolonged decelerations	2126.0	0.000159	0.000590	0.0	0.000	0.000	0.00	0.005
Abnormal Short Term Variability	2126.0	46.990122	17.192814	12.0	32.00	49.00	61.0	87.00
Mean Value of Short Term Variability	2126.0	1.332785	0.883241	0.2	0.700	1.200	1.70	7.000
Percentage of Time with Abnormal Long Term Variability	2126.0	9.846660	18.396880	0.0	0.000	0.000	11.00	91.00
Mean Value of Long Term Variability	2126.0	8.187629	5.628247	0.0	4.600	7.400	10.80	50.70
Histogram width	2126.0	70.445908	38.955693	3.0	37.00	67.50	100.0	180.0
Histogram Min	2126.0	93.579492	29.560212	50.0	67.00	93.00	120.0	159.0
Histogram Max	2126.0	164.025400	17.944183	122.0	152.0	162.0	174.0	238.0
Histogram Number of Peaks	2126.0	4.068203	2.949386	0.0	2.000	3.00	6.000	18.00
Histogram number of zeroes	2126.0	0.323612	0.706059	0.0	0.000	0.00	0.000	10.00
Histogram mode	2126.0	137.452023	16.381289	60.0	129.00	139.0	148.00	187.0
Histogram mean	2126.0	134.610536	15.593596	73.0	125.00	136.0	145.00	182.0
Histogram median	2126.0	138.090310	14.466589	77.0	129.00	139.0	148.00	186.0009
Histogram variance	2126.0	18.808090	28.977636	0.0	2.000	7.000	24.00	269.0
Histogram tendency	2126.0	0.320320	0.610829	-1.0	0.000	0.000	1.000	1.000
Fetal Health	2126.0	1.304327	0.614377	1.0	1.000	1.000	1.000	3.000

Table 1 Statistical Analysis of numerical features of the dataset

The column Count provides the number of non-empty rows in a particular feature. In this dataset, there are no empty values for all the features. The column Mean gives the mean value of the feature. The column Std gives the Standard Deviation of the feature. The column Min displays the minimum value of the feature. The columns 25%, 50%, and 75% are the percentile values of each features. This quartile information is useful in identifying outliers if any. The column Max displays the maximum value of the feature.

Checking null or missing values

It is important when analysing certain dataset to know if there are any values which are missing or have null values. The missing values are represented as NaN(Not a Number) value.

The dataset contains no null and missing values.

The figure 1 shows the bar chart for visualizing missing values.

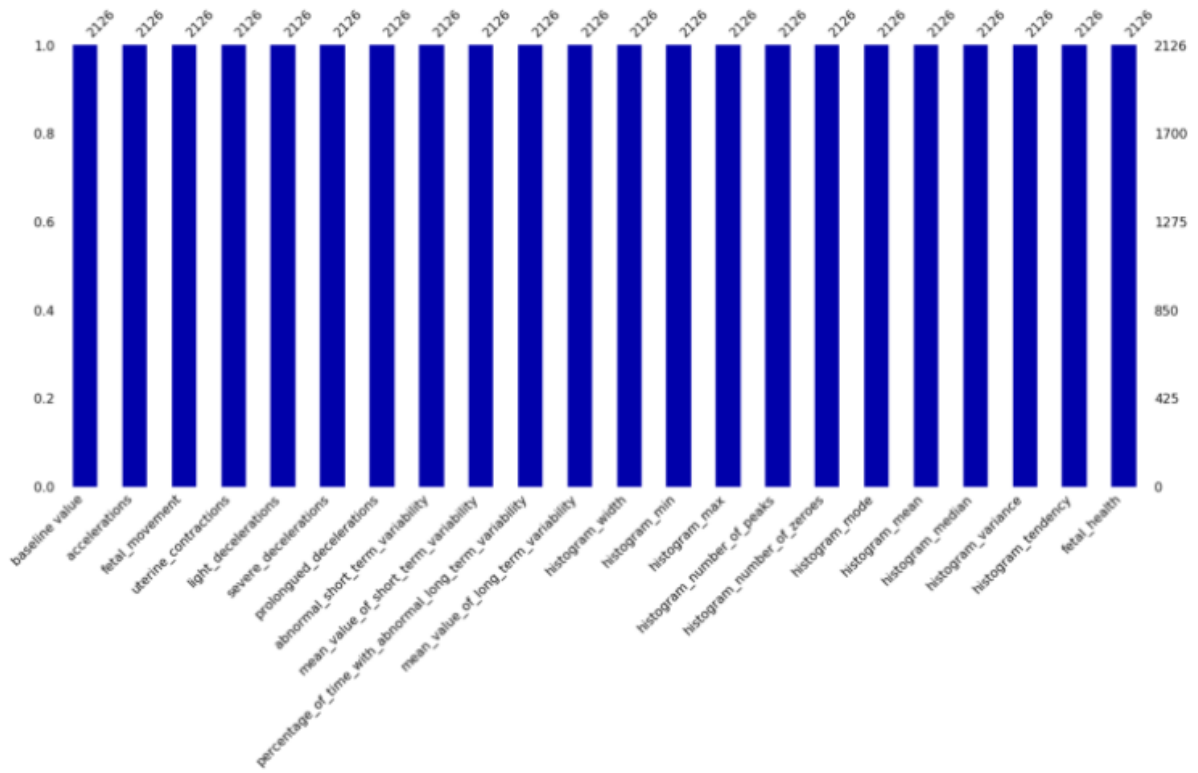


Figure 1 Bar Chart for missing value visualization

Data Visualization and Feature Analysis of dataset

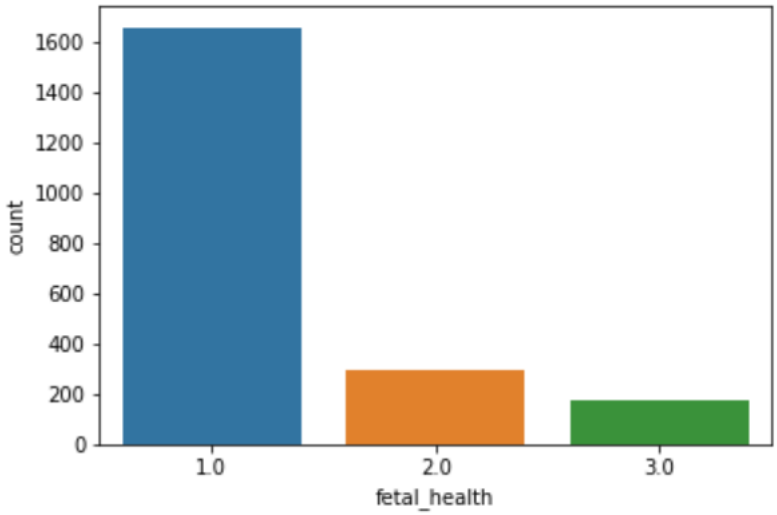


Figure 2 Fetal Health Class Count Distribution

The figure 2 shows the bar chart shows and the table 2 displays the count of each category in fetal health.

Fetal Health Category	Count
Normal	1655
Suspect	295
Pathological	176

Table 2 Fetal Health Category Count Distribution

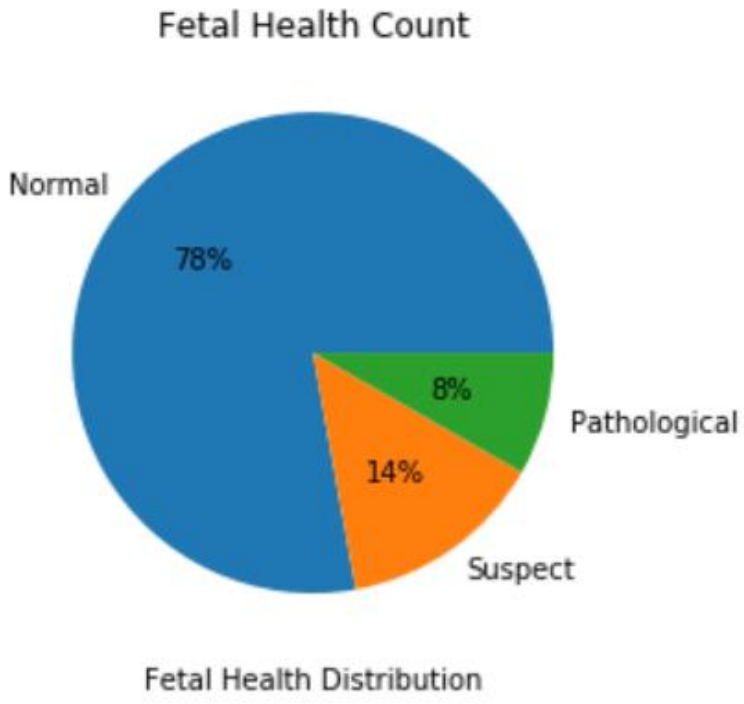


Figure 3 Fetal Health Count Pie Chart Distribution

Data visualizations of "fetal_health" column shows us the percentage of fetal health state. From the figure 3 pie chart distribution, the observation is that 78% of the total samples belong to Normal fetal health and 14% belong to Suspect and the remaining 8% belong to Pathological state.

The histogram plots below depict the distribution of the probability of the random variable near the mean value.

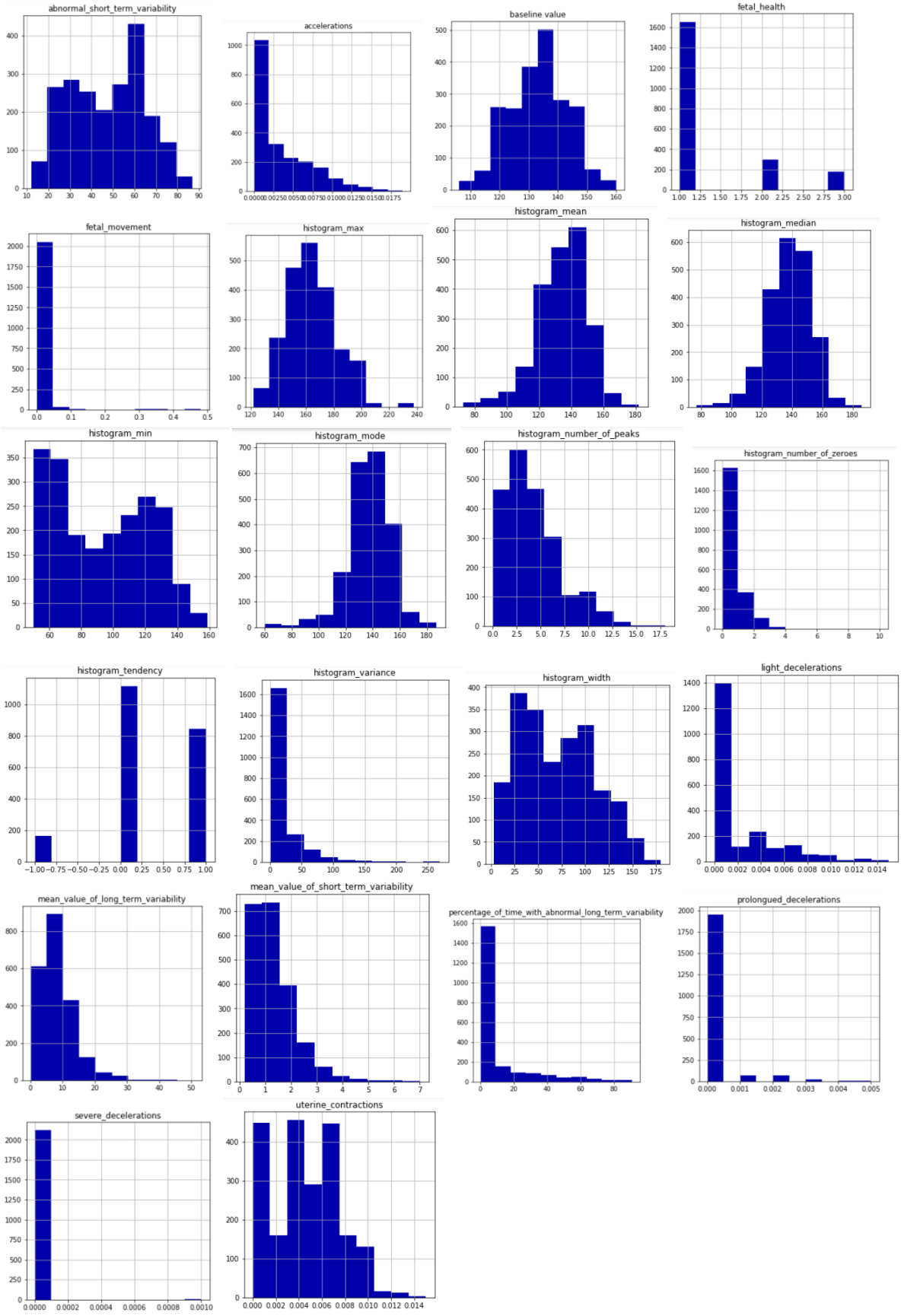


Figure 4 Histogram Plot for all 21 features

Correlation of numerical features with respect to the response output variable (Fetal Health)

The figure 5 and figure 6 represent the correlation of different numerical features in regards to the Fetal Health. This is to understand which features have the highest importance and impact the output result of the Fetal Health variable.

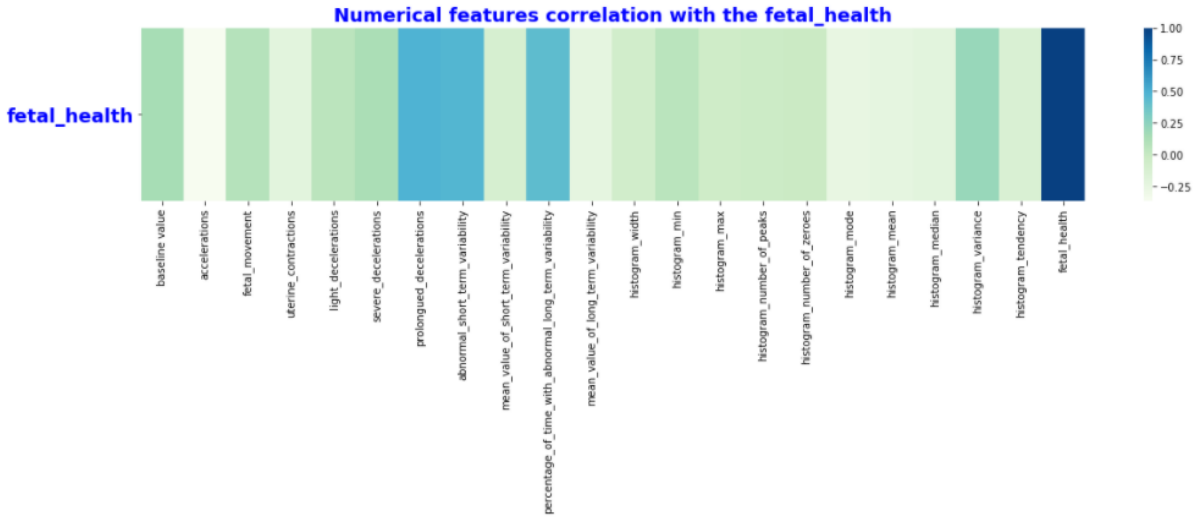


Figure 5 Numerical features correlation with fetal_health

	fetal_health
fetal_health	1
prolongued_decelerations	0.484859
abnormal_short_term_variability	0.471191
percentage_of_time_with_abnormal_long_term_variability	0.426146
histogram_variance	0.20663
baseline value	0.148151
severe_decelerations	0.131934
fetal_movement	0.08801
histogram_min	0.0631749
light_decelerations	0.0588705
histogram_number_of_zeroes	-0.0166818
histogram_number_of_peaks	-0.023666
histogram_max	-0.0452654
histogram_width	-0.0687888
mean_value_of_short_term_variability	-0.103382
histogram_tendency	-0.131976
uterine_contractions	-0.204894
histogram_median	-0.205033
mean_value_of_long_term_variability	-0.226797
histogram_mean	-0.226985

Figure 6 Correlation of features with fetal_health

The observation from the figures 5, 6 is that the features : prolongued_decelerations, abnormal_short_term_variability, percentage_of_time_with_abnormal_long_term_variability have a quite high correlation with the target variable (fetal_health).

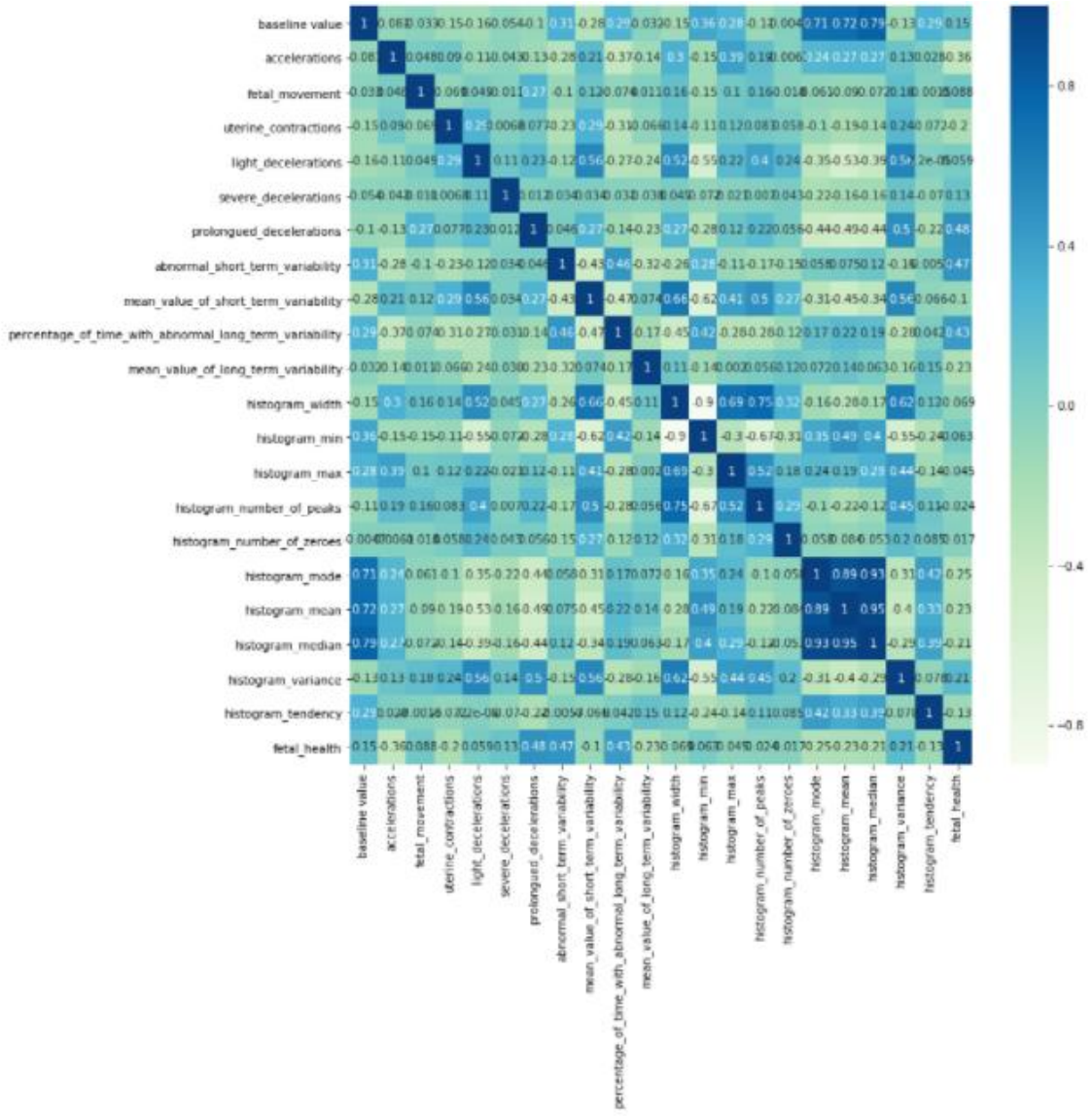


Figure 7 Correlation Heatmap of features

The figure 7 in Correlation Heat map shows the correlation between a pair of features in the Fetal Health dataset. Value closer to +1 represents highly positive correlated and values closer to -1 means highly negative correlated and values closer to 0 means there is no linear correlation between the two variables.

Observing the figure 7, the features histogram mode, histogram mean and histogram median are highly positive correlated to each other.

Principal Component Analysis

PCA helps to reduce the dimensionality of large data sets into smaller data sets that can still tell most of the information and variable in the original dataset.[3]

	Mean	Variance	Explained Variance	Explained Variance Ratio
Baseline Value	133.303857	96.84221570	6.06216677	0.27542342
Accelerations	0.003178	0.00001494	3.66275993	0.16641078
Fetal_movement	0.009481	0.00217770	2.29934388	0.10446647
Uterine Contractions	0.004366	0.00000868	1.50697880	0.06846682
Light decelerations	0.001889	0.00000876	1.23276004	0.05600819
Severe decelerations	0.000003	0.00000000	1.02752108	0.04668353
Prolonged decelerations	0.000159	0.00000035	0.99228402	0.04508260
Abnormal Short Term Variability	46.990122	295.59284356	0.92700998	0.04211700
Mean value of short term variability	1.332785	0.78011525	0.76458835	0.03473767
Percentage of Time with Abnormal Long Term Variability	9.846660	338.44518178	0.66587920	0.03025300
Mean Value of Long Term Variability	8.187629	31.67715984	0.57786386	0.02625418
Histogram width	70.445908	1517.54601383	0.52516566	0.02385994
Histogram min	93.579492	873.80614864	0.39623194	0.01800207
Histogram max	164.025400	321.99370749	0.34404079	0.01563086
Histogram number of peaks	4.068203	8.69887555	0.26682656	0.01212278
Histogram number of zeroes	0.323612	0.49851984	0.25993769	0.01180979
Histogram mode	137.452023	268.34663826	0.17618088	0.00800446
Histogram mean	134.610536	243.16024658	0.13067092	0.00593679
Histogram median	138.090310	209.28219313	0.11608239	0.00527399
Histogram variance	18.808090	839.70338863	0.04932508	0.00224099
Histogram tendency	0.320320	0.37311162	0.02673513	0.00121466
Fetal health	1.304327	0.37745891	0.00000000	0.00000000

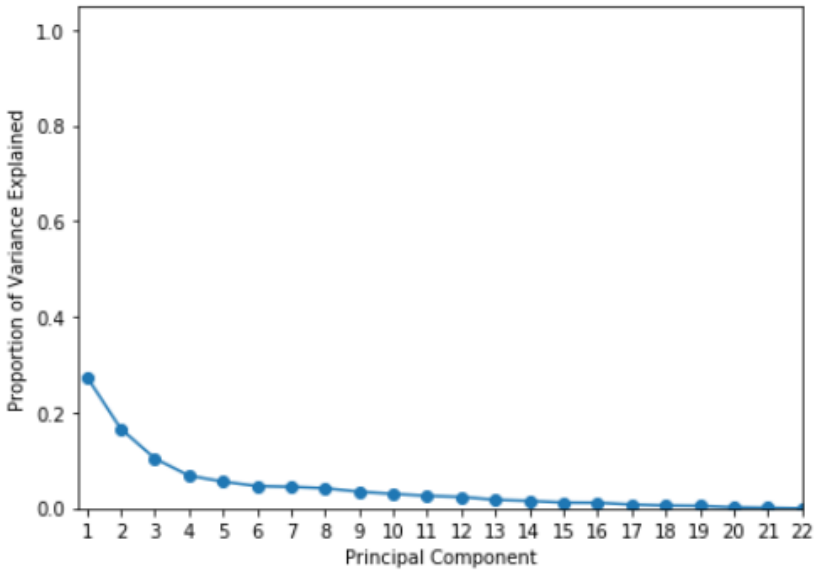


Figure 8 Principal Component Vs Proportion of Variance explained

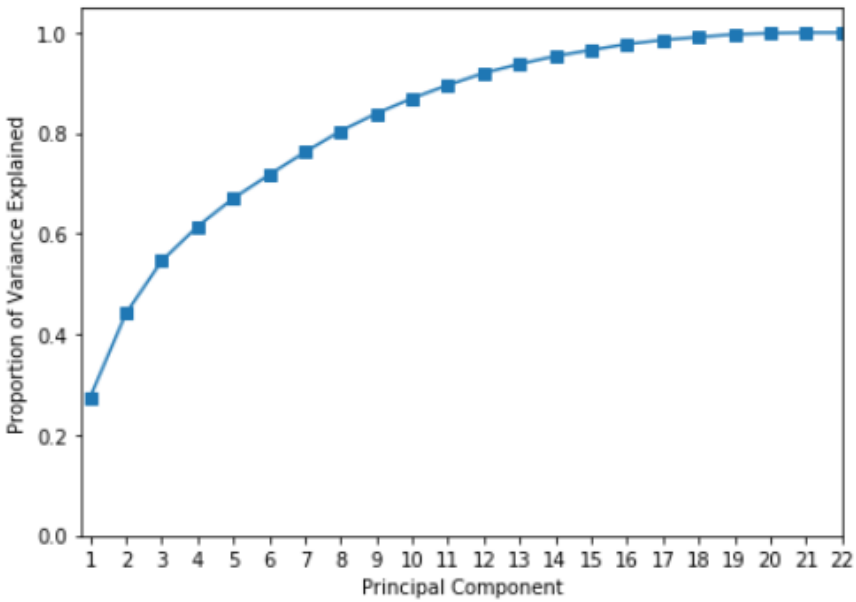


Figure 9 Principal Component Vs Proportion of cumulative Variance explained

From the explained variance and explained variance ratio of each of the PCA components representing each feature, the features baseline value, accelerations, fetal movement, uterine contractions, light decelerations, severe decelerations, prolonged decelerations, abnormal short term variability, these features are most important in determining the target class variable as they have high explained variance and high explained variance ratio.

Dataset Analysis Methodology

The below steps also illustrated in Figure 10 are followed for implementing the various classification model algorithms as illustrated in the figure.

- The input data is collected, visualized and analysed.
- The features which are most important and relevant in determining the target class are selected using different feature selection methods.
- The dataset dimension reduction is done by eliminating the least significant features
- Feature Scaling is performed using standardization.
- The dataset is splitted into two sets training sets consisting of 70% and testing set consisting of 30%.
- Different classification algorithms are applied on the dataset.
- Finally, the model performance is evaluated by various evaluation metrics parameters.

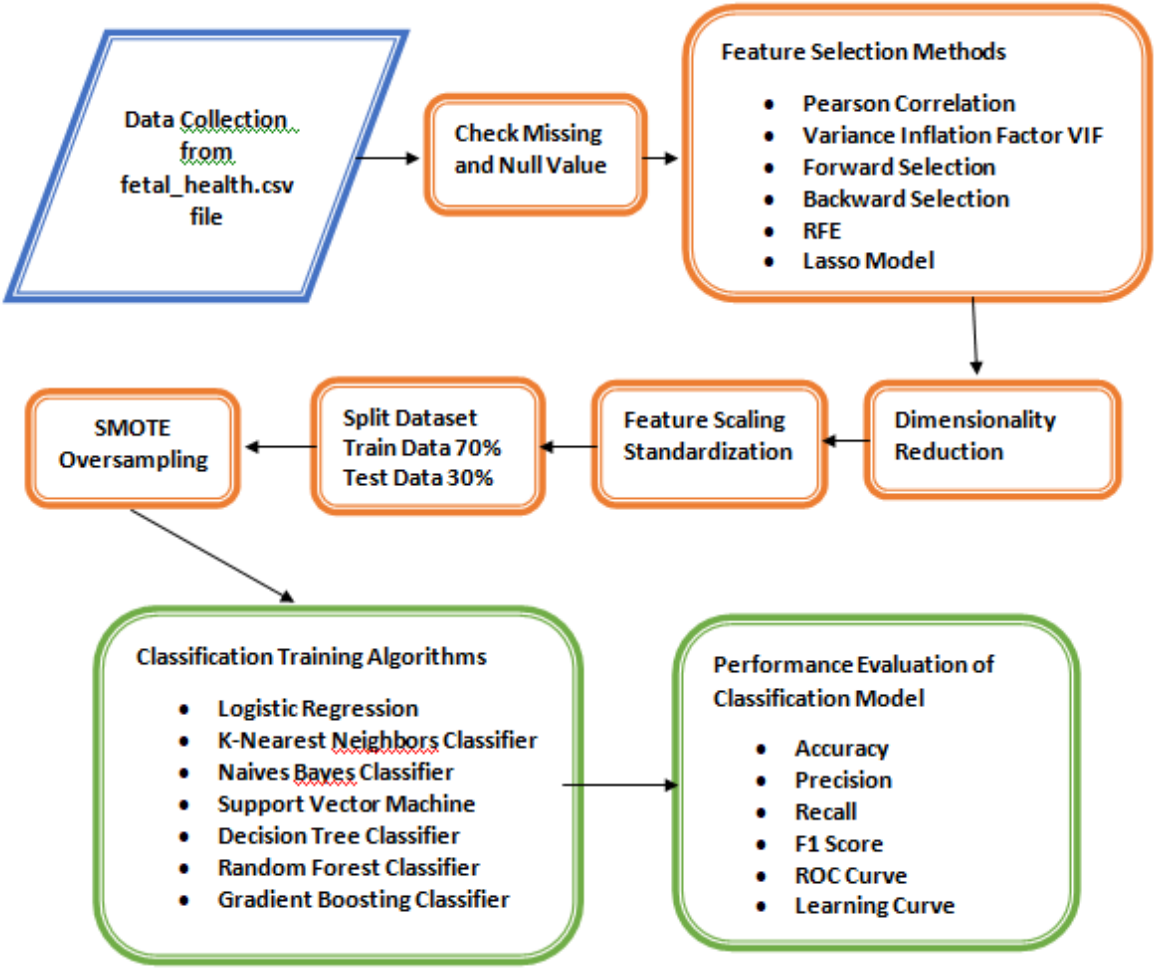


Figure 10 Dataset Analysis Methodology

Pre-processing the Dataset

Selecting the Features set for generating predictive model. There are multiple methods to select the features which are most important in determining the target class such as Filter methods, Wrapper methods and Embedded methods.

The following feature selection methods are applied on the dataset.

- Filter Methods: Pearson Correlation method:
- Wrapper Methods: Enter Method VIF, Forward Selection Method, Backward Elimination Method, RFE Method
- Embedded Method – Lasso Model

Feature	Pearson Correlation	Enter Method VIF	Forward Selection	Backward Elimination	RFE	Lasso Model	Total Count
Baseline Value	False	True	True	True	True	True	5
Accelerations	True	False	True	False	True	False	3

Fetal Movement	False	False	False	False	True	False	1
Uterine Contractions	True	False	True	True	True	False	4
Light decelerations	False	False	False	False	True	False	1
Severe decelerations	False	False	True	True	True	False	3
Prolonged decelerations	True	False	True	True	True	False	4
Abnormal Short Term Variability	True	True	True	True	True	True	6
Mean Value of Short Term Variability	False	True	False	False	False	False	1
Percentage of Time with Abnormal Long Term Variability	True	False	True	True	True	True	5
Mean Value of Long Term Variability	True	True	False	True	False	True	4
Histogram width	False	False	False	True	False	False	1
Histogram Min	False	False	True	False	False	True	2
Histogram Max	False	False	False	True	False	True	2
Histogram Number of Peaks	False	True	False	False	False	False	1
Histogram number of zeroes	False	False	False	False	True	False	1
Histogram mode	True	True	True	False	False	True	4
Histogram mean	True	True	False	False	False	True	3
Histogram median	True	True	True	True	True	True	6
Histogram variance	True	False	True	True	False	True	4
Histogram tendency	False	False	False	False	True	False	1

Table 3 Selected Features in different Feature Selection Techniques

The table 3 displays whether a feature is selected in all or most of the methods. The total column indicates the number of times the feature is being selected in the different methods. In

this case, based on the total value column, it is evident that the features abnormal_short_term_variability, histogram_median, baseline_value, percentage of time with abnormal long term variability are crucial and important attributes in determining the fetal health whereas fetal movement, light decelerations, mean value of short term variability, histogram width, histogram number of peaks, histogram number of zeroes and histogram tendency width are least important attributes.

The feature selection is done on the basis of the condition where the feature is picked for at least 3 times among all the above feature selection techniques. Below are the final selected features as response variables for creating predictive models.

- Baseline Value
- Accelerations
- Uterine Contractions
- Severe decelerations
- Prolonged decelerations
- Abnormal Short Term Variability
- Percentage of Time with Abnormal Long Term Variability
- Mean Value of Long Term Variability
- Histogram mode
- Histogram median
- Histogram variance
- Fetal Health

Feature Scaling

This is required to ensure that all the variable values are in the same range or in same scale so that there are no dominant features than others. For this dataset, standardization is applied such that the mean of the distribution is 0 and standard deviation is 1. [4]

$x' = (x - \bar{x}) / \sigma$ where $\bar{x}$ is mean and σ is variance

Train Test Split

The dataset is splitted into train dataset and test dataset with 70% and 30% ratio from the given dataset. The train dataset would be used for training the different models and the unknown test dataset would be used for testing the model. This ensures better calculation of performance of the model.

SMOTE Oversampling

Synthetic minority oversampling technique (SMOTE) is useful where there is less data about the minority class for learning the decision boundary. This technique oversamples the data in the minority target class. [5]

Building Machine Learning Models:

The goal of for this dataset analysis is to predict Fetal Health with least prediction error.

The prediction of the fetal health into classes as Normal, Suspect and Pathological is a multiclass classification problem where the output variable which is fetal health in this case can be classified into any one out of the three classes.

In the classification process, the training dataset is analysed to identify the various boundary conditions in order to determine each target class. After knowing the boundary conditions a model is designed to predict the target class as accurate as possible. [17]

The classification techniques has the terminologies such as classifiers, classification model, feature, binary classification, multiclass classification and multilabel classification. A Classifier is an algorithm that maps the response features to a particular category. The Classification model draw some conclusions from the input values given for training[22]. It will predict the class labels/categories for the new data. Feature is an individual measurable property of a phenomenon being observed.[6][17]

A Binary Classification task is having two outcomes as the name suggests. Multi-class classification task is to classify the input data into only one of the many target classes. In case of multi-label classification, the task is to map the sample data to a set of one or more target classes. [17]

The following are different types of classification algorithms that can be applied on a dataset for addressing a classification problem: [17]

- Linear Classifiers: Logistic Regression, Naïve Bayes Classifier, Fisher's linear discriminant
- Support Vector Machines: Support Vector Classifier
- Kernel Estimation: K nearest neighbor
- Decision Trees: Random Forests
- Quadratic Classifiers
- Neural Networks
- Boosting Technique: Gradient Boosting, Adaptive Boosting
- Linear Vector Quantization

For the CTG dataset for classifying the fetal state, the below classification algorithms are modelled and evaluated to build the predictive model with highest possible accuracy.

Logistic regression- This algorithm would fit the data to a logit function and then predict the probability of occurrence of the event based on the input variables which are independent. In this case it would predict the given data into Normal, Suspect or Pathological [17]

K-Nearest neighbors –It is a supervised learning algorithm used for classification problem that assumes similar items exist close to each other. [8]

Naive Bayes Classifier- This is based on Bayes' theorem where all the predictor features are independent. It is suitable for huge datasets and multiclass classification problems which also applied to the CTG dataset. [17]

Support vector machines – This is a supervised machine learning algorithm mainly used for classification problems. Here, each tuple data is plotted as a point in n dimensional space where n denotes number of features or support vectors. A classification is obtained by identifying a hyperplane which distinguishes into different classes. [14]

Decision trees – It is a predictive model approach by building a decision tree to determine the target label class of the given data. [9]

Random forest – It is an ensemble learning classification method operating by constructing several decision trees and calculating classification. [10]

Gradient Boosting It is machine learning algorithm for solving classification problems, producing ensemble of weak prediction model that is decision trees building models in stage wise pattern. [11]

Performance Evaluation Metrics

After training the model the most important part is to evaluate the classifier to verify its applicability. Below methodologies and parameters are used to assess the model [12]

Holdout method **–** In this method, the available dataset is divided into two groups namely train and test data which comprises of 70% and 30% of data respectively. The train data is used for training the model and the remaining test data is used for determining the correctness of prediction. [12]

Cross-validation- K-fold cross validation is a evaluation method to ensure that the model is not over-fitted. The dataset is partitioned into k mutually exclusive subsets randomly almost having same size and out of them one is picked up for testing and others for utilized for training. The entire process is repeated for k folds. This process is iterated throughout the whole k folds [12]

	Positive	Negative
Predicted as positive	TP (True Positive)	FP (False Positive)
Predicted as negative	FN (False Negative)	TN (True Negative)

Table 4 Confusion Matrix

	Calculation Formula	**Evaluation description**
Accuracy	(TP+TN)/(TP+FP+FN+TN)	Overall efficiency of classifier
Se(Recall)	TP/(TP+FN)	Efficiciency of classifer to categorize positively labelled data
Sp	TN/(TN+FP)	Efficicieny of classifier to categorize negatively labelled data

Precision	TP/(TP+FP)	Data with positive labels correctly classified by the classifer

Table 5 Performance Evaluation Metrics calculation formula[15]

Precision It is also known as positive prediction value which is fraction of relevant instances among retrieved instances.[15]

Recall –It is also known as sensitivity is the fraction of relevant instances among those that were actually retrieved.[15]

Confusion Matrix - It displays the classification model prediction results. It helps to understand how far is the model is correct and insights into type of errors. [13]

ROC curve (Receiver Operating Characteristics) – ROC curve is a visual representation for comparing the classification model that displays the graph of between the true positive rate and false positive rate. The area under the ROC curve gives the accuracy of the model. The more the model is away from the diagonal, the more it is accurate. An area of 1.0 denotes that the model is having perfect accuracy. [12]

Learning Curve These plots help to show developmental changes in performance during learning. It is also used to detect based on the train and test/validation dataset as under-fitted, over-fitted or well-fitted model. [18]

RESULTS

The different classifier algorithms were performed on the reduced imbalanced dataset.

The figure 11 shows the distribution of the sample count between the different classes Normal, Suspect and Pathological in both training dataset and test dataset before applying SMOTE oversampling. Clearly, the samples in the fetal_health class Normal is fairly large when compared to other classes Suspect and Pathological in both the Train and Test Dataset

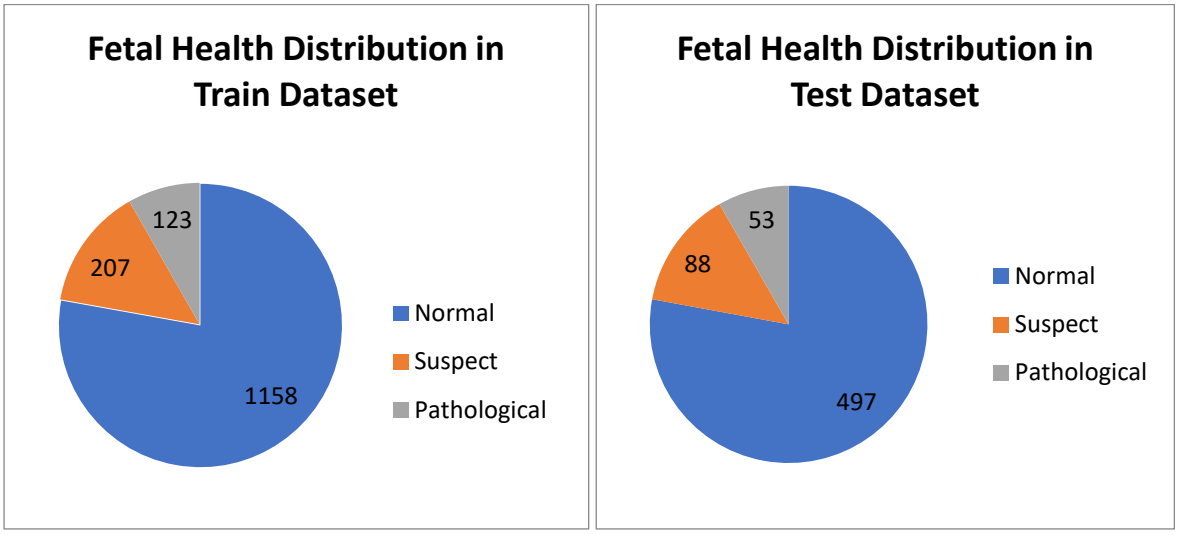


Figure 11: Fetal Health Distribution in Train and Test Dataset before oversampling

Before applying the SMOTE oversampling technique and after applying SMOTE oversampling, following were the comparative results that were obtained.

	Before SMOTE oversampling			After SMOTE oversampling		
	Accuracy Score	Train Score	Test Score	Accuracy Score	Train Score	Test Score
Logistic Regression	0.893	0.901	0.893	0.839	0.889	0.839
K Nearest Neighbour	0.9	0.951	0.9	0.91	0.975	0.91
Random Forest	0.923	0.999	0.923	0.93	1.0	0.93
Decision Trees	0.908	0.999	0.908	0.904	1.0	0.904
Naives Bayes Algorithm	0.826	0.857	0.826	0.746	0.792	0.746
SVM	0.893	0.929	0.893	0.887	0.939	0.887
Gradient Boosting	0.929	0.989	0.929	0.934	0.992	0.934

Table 6 Accuracy, Train and Test Scores for models before/after SMOTE oversampling

The figure 12 shows the distribution of the fetal health class in both train and test dataset after SMOTE oversampling. The samples from the minority classes are oversampled to obtain equal distribution in the samples among all the three classes in both Train and Test dataset.

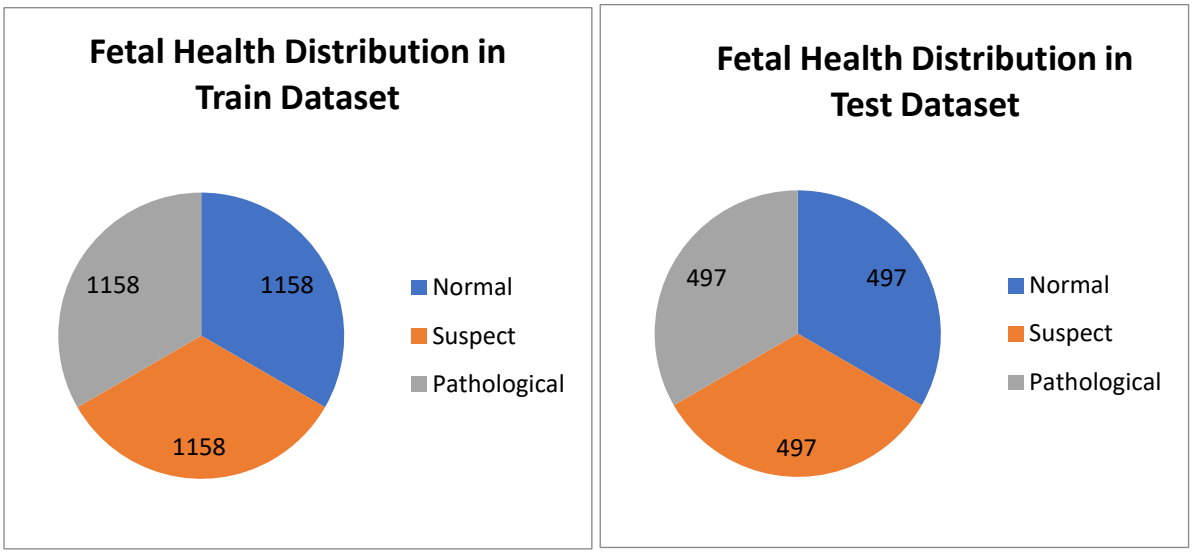


Figure 12: Fetal Health Distribution in Train and Test Dataset after oversampling

After applying SMOTE oversampling technique, below were the results obtained for all the three cases: Before Hyperparameter tuning, Hyperparameter tuning using GridSearchCV and hyperparameter tuning using RandomizedSearchCV.

	After SMOTE Oversampling							
	Before Hyperparameter Tuning		After Hyperparameter Tuning using Grid Search			After Hyperparameter Tuning using Randomized Search		
	Accuracy Score	Best Score	Accuracy Score	Best Score	Time Taken	Accuracy Score	Best Score	Time Taken
Logistic Regression	0.839	0.889	0.838	0.887	1.1 min	0.838	0.887	10.9 sec
K Nearest	0.886	0.975	0.875	0.977	11.2	0.875	0.977	11.3

Neighbour					min			sec
Random Forest	0.93	1.0	0.932	0.981	11.6 min	0.933	0.97	2 min
Decision Trees	0.917	1.0	0.858	0.962	-	0.859	0.929	-
Naives Bayes Algorithm	0.746	0.792	0.783	0.817	13.3 sec	0.775	0.815	1.1 sec
SVM	0.887	0.939	0.863	0.973	3.8 min	0.84	0.97	10 sec
Gradient Boosting Machine	0.934	0.992	0.93	0.979	19.8 min	0.93	0.977	5.4 min

Table 7 Scores before/after Hyperparameter Tuning

The observation from SMOTE oversampling technique application is that it significantly improved the accuracy of K-Nearest Neighbors, Random Forest and Gradient Boosting models. Since this dataset is imbalanced therefore it is important to use SMOTE technique to achieve better results to avoid any bias towards dominant class.

Another interesting fact is that the accuracy of the model also increased after applying Hyperparameter Tuning in Naives Bayes and Random Forest algorithm. For the algorithms K-Nearest Neighbors, Naives Bayes, SVM the best score is improved.

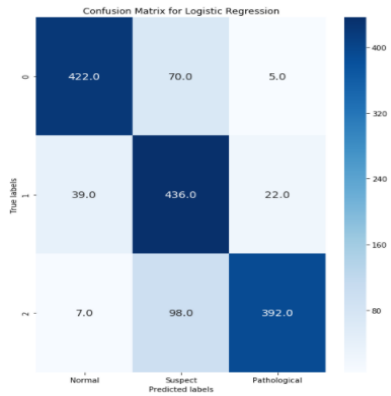
When compared to both gridsearchCV and RandomizedSearchCV hyperparameter tuning methods, GridSearchCV provided better results with slightly more accuracy and best score when compared to RandomizedSearchCV in most of the classifier models, but it took longer execution time.

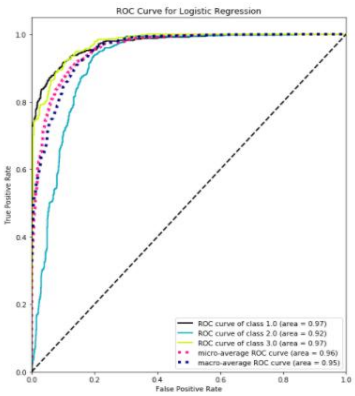
The table 8 shows performance metrics were calculated for all the classification models after applying Hyperparameter Tuning method.

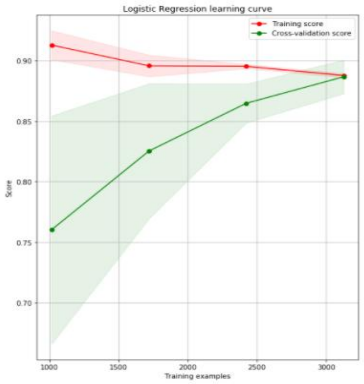
	Accuracy	**RMSE**	**Precision**			**Recall**			**F1 score**		
			N	S	P	N	S	P	N	S	P
Logistic Regression	0.84	0.431	0.90	0.72	0.94	0.85	0.88	0.79	0.87	0.79	0.86
K Nearest Neighbour	0.88	0.383	0.81	0.87	0.96	0.96	0.76	0.90	0.88	0.81	0.93
Random Forest	0.93	0.29	0.89	0.94	0.97	0.96	0.87	0.97	0.92	0.91	0.97
Naives Bayes Algorithm	0.78	0.528	0.91	0.65	0.89	0.82	0.90	0.63	0.86	0.75	0.73
Decision Trees	0.86	0.441	0.83	0.81	0.93	0.91	0.81	0.85	0.87	0.81	0.89
SVM	0.86	0.506	0.77	0.88	0.98	0.95	0.82	0.82	0.85	0.85	0.89
Gradient Boosting Machine	0.93	0.296	0.88	0.95	0.97	0.97	0.86	0.97	0.92	0.90	0.97

Table 8 Performance Metrics for classification models

Logistic Regression Model:

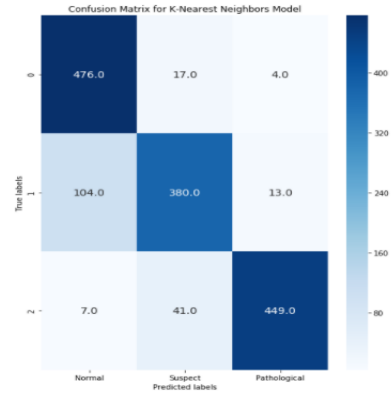


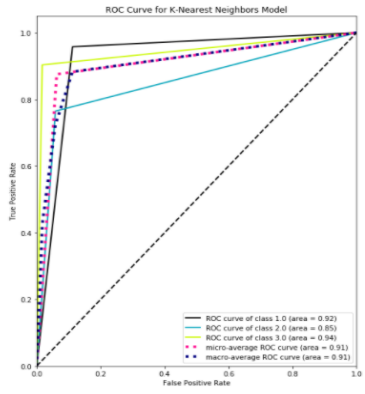


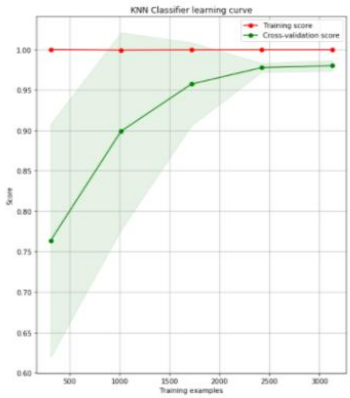


In case of Logistic Regression Model, from the confusion matrix the maximum correctly predicted class labels were Suspect fetal health category with 436 out of 497, followed by Normal with 422 out of 497 and Pathological predicting 392 tuples out of 497. Also, ROC Curve shows that the ROC curve area for the class 1.0 and class 3.0 is 0.97 which is more than the ROC curve area of class 2.0 which is 0.92.

K-Nearest Neighbors Model

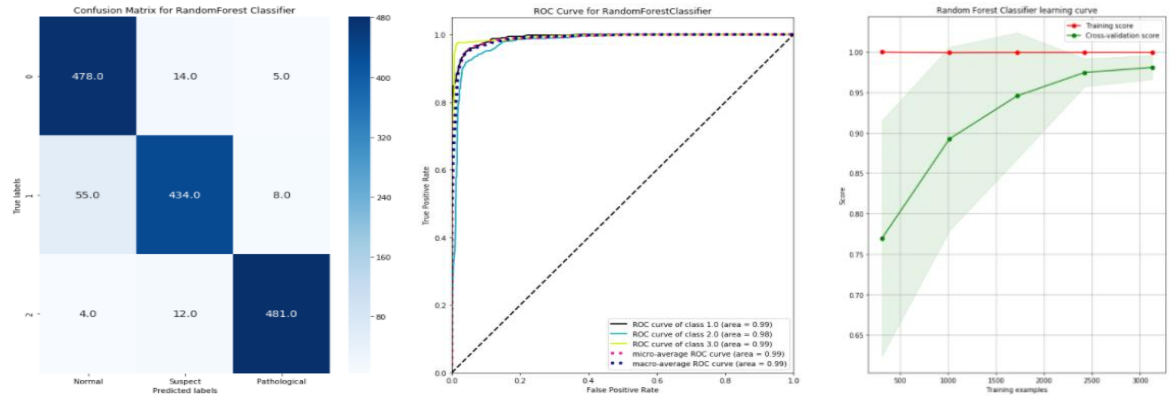






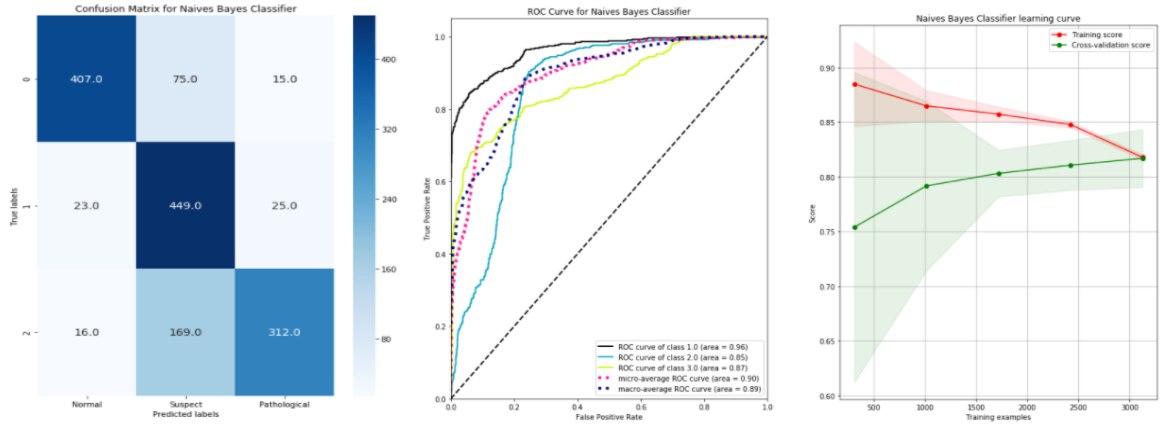
K-Nearest Neighbors Model correctly predicted Normal class labels as highest with 476 out of 497 followed by Pathological 449 and the least being Suspect 380. Also, ROC Curve area for class 3.0 is 0.94 which is more than the ROC curve area of class 1.0 Normal which is 0.92 and class 2.0 Suspect as 0.85.

Random Forest Classifier Model



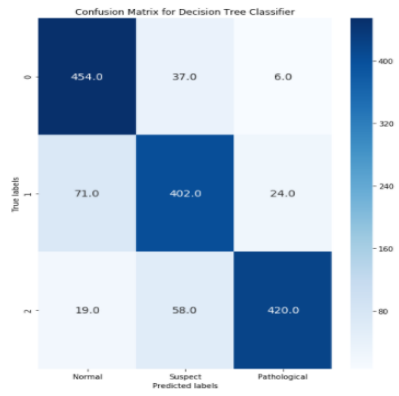
Random Forest Model was able to correctly predict class 1 Normal (478 out of 497 samples) and class 3 Pathological (481 out of 497 samples) better when compared to class 2 Suspect (434 out of 497 samples). The ROC curve area for class 1 and class 3 is 0.99 which is slightly more than class 2 which is 0.98

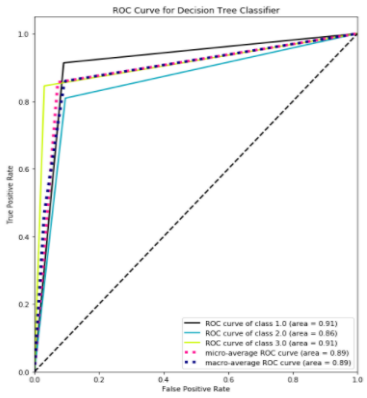
Naives Bayes Classifier Model

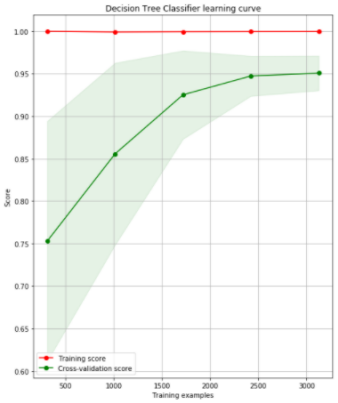


Naives Bayes Model was able to correctly predict class 2 (449 out of 497 samples) Suspect comparatively better than class 1 normal and class 3 pathological. The ROC curve area for class 1 is larger 0.96 when compared to class 2 which is 0.85 and class 3 which is 0.87.

Decision Tree Classifier

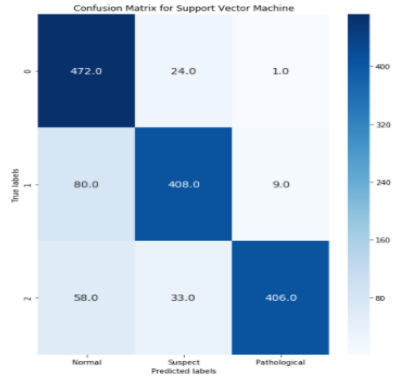


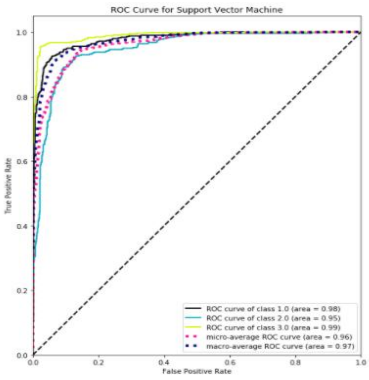


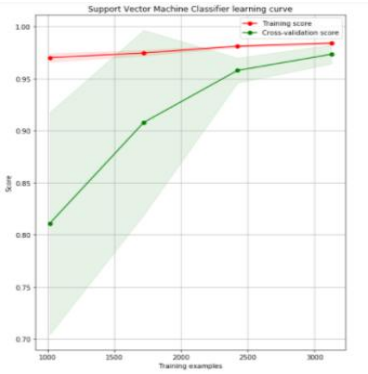


Decision tree classifier was able to correctly predict more samples from the class 1 Normal when compared to class 2 suspect and class 3 pathological. The ROC curve area of class 1 and class 3 is 0.91 when compared to that of class 2 which is 0.86

Support Vector Machine

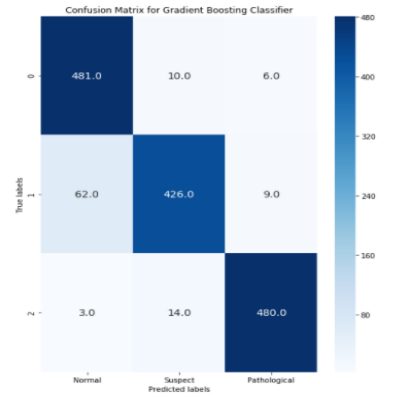


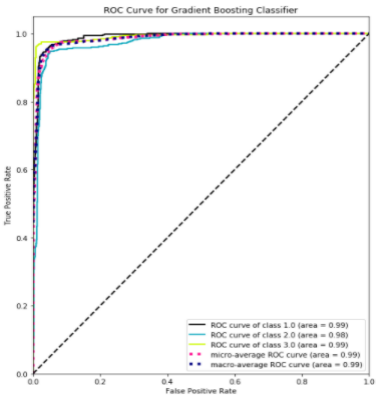


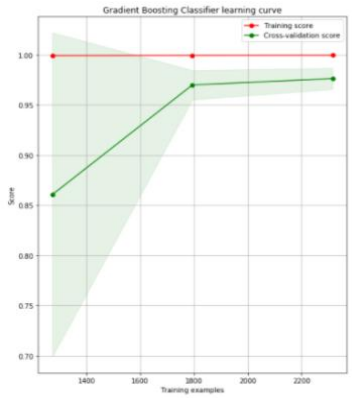


Support Vector Machine predicted class 1 normal samples the most correct with 472 out of 497 when compared to class 2 suspect and class 3 pathological. Also, the ROC curve area for class 1, class 2, class 3 are 0.98, 0.95 and 0.99 respectively.

Gradient Boosting Classifier







Gradient Boosting classifier was able to correctly predict more samples for both the class 1 Normal (481 out of 497) and class 3 pathological (480 out of 497). It also predicted 426 out

of 497 samples correctly for class 2 Suspect. The ROC curve area for class 1 , class 2 and class 3 are 0.99, 0.98 and 0.99 respectively.

DISCUSSION

The table 8 gives the performance metrics of all the classification algorithms. For better performance evaluation of the different classification models, 10 fold cross validation was implemented. The Random Forest Model and the Gradient boosting model gave the highest accuracy of 93%. Also, the values for the root mean squared error RMSE for both Random Forest and Gradient Boosting model are 0.29 and 0.296 which indicates that both the models could more accurately predict the response variable which is fetal health in this case and also are well fitted since the value of RMSE is less when compared to other models.

The next best model was K-Nearest Neighbor with 88% accuracy and 0.383 RMSE value. The models decision tree and Support vector machine had 86% accuracy and Logistic regression with accuracy of 84% but with high RMSE value which means that models are not well-fitted and also cannot predict the target class more accurately when compared to Random Forest, Gradient Boost and K-Nearest Neighbor models. Naives Bayes model was least effective for this data set with the accuracy of 78% and RMSE value of 0.528.

Also, on the basis of the confusion matrix results for Random Forest and Gradient Boosting models, there were many interesting data insights for each of the fetal health classes. Both the random forest and gradient boosting models best predicted the normal class with gradient boosting model predicting slightly better than the random forest. Even, K-Nearest Neighbors and Support Vector machine were only little behind in predicting the normal class accurately compared to the earlier best prediction models for Normal class.

Naives Bayes model was most successful among all the models in best predicting the suspect class. Pathological class was best predicted by both Gradient boosting model and Random Forest models.

The highest ROC curve area belonged to Gradient Boosting and Random Forest models (ROC=0.99) which proves that these models are suitable to correctly classify and predict the fetal health based on the input data.

CONCLUSION

The Cardiotocography CTG medical data is helpul for medical practitioners in identifying fetal health issues and do medical intervention in critical cases preventing any damage on the baby and expectant mother. Therefore, the identification of fetal abnormalities by utilitizing CTG dataset, a decision support system built on using different machine learning models could be of great medical aid.

With the experimental results on the dataset, the gradient boosting and random forest models are best suited models which could be utilized for predicting and classifying the CTG data into different fetal health classes.

For further assessing the performance, the models need to be tested on larger dataset of CTG data. This would ensure the authencity of the achieved results and could also improve the accuracy and performance further.

Future scope also includes testing the fetal health CTG dataset with more complex ensemble classification algorithms such as bagging, boosting and stacking.

REFERENCES:

[1] https://www.kaggle.com/andrewmvd/fetal-health-classification

Ayres de Campos et al. (2000) SisPorto 2.0 A Program for Automated Analysis of Cardiotocograms. J Matern Fetal Med 5:311-318

[2] Repository UIML. [Last accessed on 2018 Jun 11]. Available from: https://archive.ics.uci.edu/ml/index.php .

[3] https://en.wikipedia.org/wiki/Principal_component_analysis

[4] https://medium.com/@rahul77349/feature-scaling-why-it-is-required-8a93df1af310

[5] https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/

[6] https://machinelearningmastery.com/types-of-classification-in-machine-learning/

[7] https://www.edureka.co/blog/classification-algorithms/

[8] https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

[9] https://medium.com/swlh/decision-tree-classification-de64fc4d5aac

[10] https://en.wikipedia.org/wiki/Random_forest

[11] https://en.wikipedia.org/wiki/Gradient_boosting

[12] https://towardsdatascience.com/machine-learning-classifiers-a5cc4e1b0623

[13] https://machinelearningmastery.com/confusion-matrix-machine-learning/#:~:text=A%20confusion%20matrix%20is%20a,two%20classes%20in%20your%20dataset.

[14] https://engmrk.com/module-16-support-vector-machine-2/?utm_campaign=News&utm_medium=Community&utm_source=DataCamp.com

[15] https://en.wikipedia.org/wiki/Precision_and_recall#:~:text=In%20pattern%20recognition%2C%20information%20retrieval,of%20relevant%20instances%20that%20were

[16] https://www.researchgate.net/publication/299387264_Classification_of_cardiotocograms_using_random_forest_classifier_and_selection_of_important_features_from_cardiotocogram_signal

[17] https://dzone.com/articles/introduction-to-classification-algorithms

[18] https://machinelearningmastery.com/learning-curves-for-diagnosing-machine-learning-model-performance/#:~:text=Learning%20curves%20are%20plots%20that,%2C%20or%20well%2Dfit%20model.

[19] http://www.ncbi.nlm.nih.gov/pubmed/16856111

[20] https://indicators.report/targets/3-2/

[21] https://www.researchgate.net/profile/Muhammad_Arif32/publication/299387264_Classification_of_cardiotocograms_using_random_forest_classifier_and_selection_of_important_features_from_cardiotocogram_signal/links/58da1f7e92851ce5e92bb189/Classification-of-cardiotocograms-using-random-forest-classifier-and-selection-of-important-features-from-cardiotocogram-signal.pdf

[22] https://analyticsindiamag.com/7-types-classification-algorithms/

APPENDIX

Fetal Health dataset link

https://www.kaggle.com/andrewmvd/fetal-health-classification

fetal_health.csv file