Abstract:

The Synthetic Infant Health Classification dataset is used in the medical field to predict outcomes using generated data rather than private data. It is also referred to as mimic data. Medical data is often kept confidential, which can hinder IT innovations in the industry. Therefore, mimic data is designed to make predictions and data analysis, which helps to prevent privacy breaches. Developing mimic data also helps organizations to broaden the use of the data, increasing the number of users and contributors who can analyze the data. Most datasets containing health information are not readily available for use because they contain confidential information about individuals. Identifiable records can't also be easily shared as organizations need to comply with certain regulations. Researchers and analysts continue to face many barriers when accessing essential datasets. Therefore, synthetic datasets are preferred over any official dataset. Infants aged 0-30 days have extremely sensitive health because of this another barrier gets created for their health data analysis and study. However, it is extremely important to analyze their health issues and prepare good models that will help in the early diagnosis of diseases in infants. Early diagnosis of these diseases will not only help doctors but also many industries on these lines. Machine Learning and Data Science have very wide applications one of which is this. Due to the advantages of Machine Learning and Data Science, the process of diagnosis speeds up and helps doctors to take precautions accordingly. In this article, I am using my knowledge of Machine Learning and Data Science to predict the sickness of babies. I am utilizing Python libraries such as Seaborn, Pandas, Numpy, and Sklearn to comprehend the data. For my Data Science and Machine learning project along with these libraries, I have used the Random Forest algorithm and K nearest Neighbor algorithm to predict the sickness of the infant babies. The data provided in the dataset is used to train the model.

Introduction :

In the medical field data plays a huge role. it is vital to analyze from time to time. Therefore synthetic data is gaining more importance these days which is very similar to actual data but still abstracts the data properly. The data is realistic hence it gets easier for the data analyzers to make realistic predictions which boosts the innovations in the medical field. The dataset used in this project is taken from (https://mitu.co.in/dataset).

The dataset includes various details about the infants, the columns titles are BirthAsphyxia, HypDistrib, HypoxiaInO2, CO2, ChestXray, Grunting, LVHreport, LowerBodyO2, RUQO2, CO2Report, XrayReport, Disease, GruntingReport, Age, LVH, LungParench, LungFlow, and Sick.

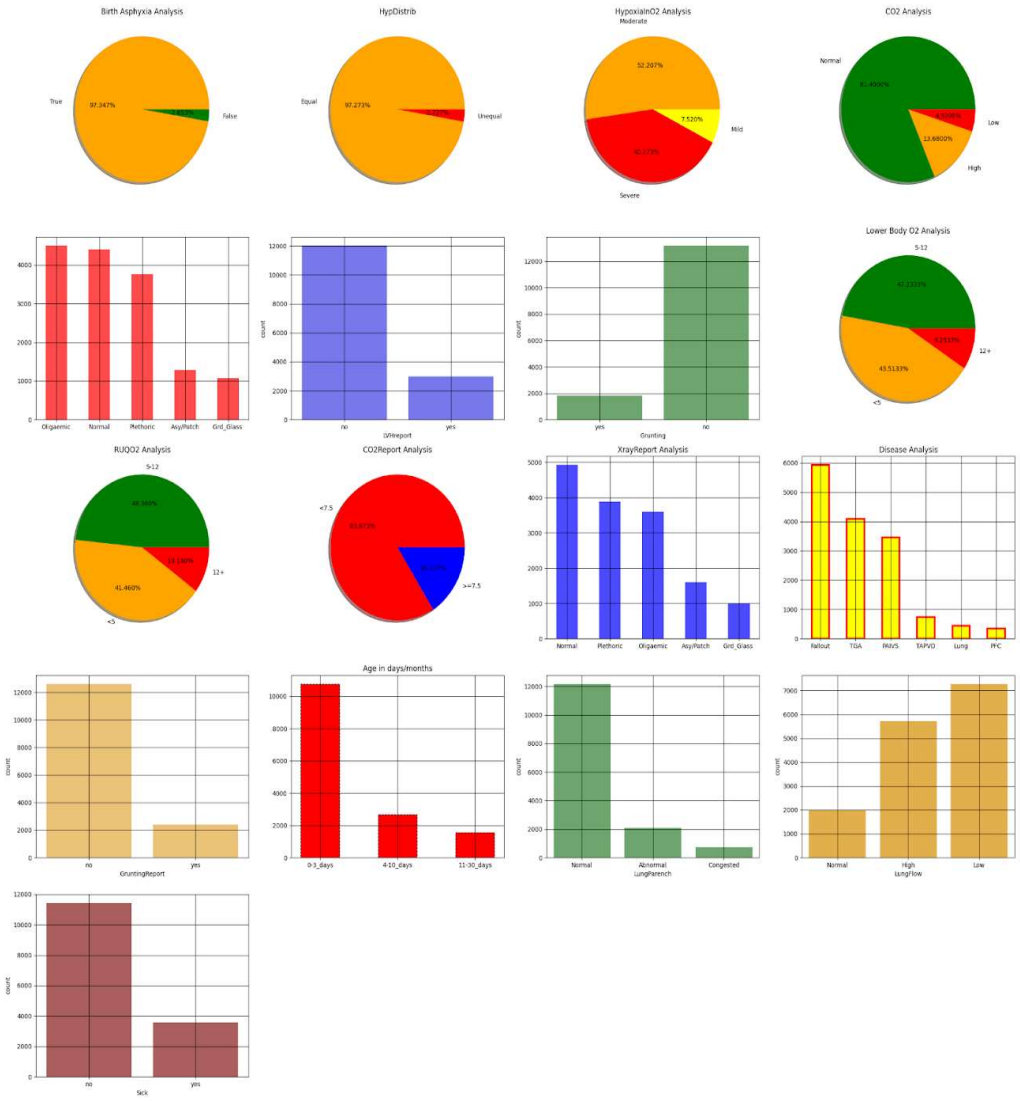
	Unnamed: 0	BirthAsphyxia	HypDistrib	HypoxiaInO2	CO2	ChestXray	Grunting	LVHreport	LowerBodyO2	RUQO2	...	XrayReport	Disease	GruntingReport
0	0	no	Equal	Severe	Normal	Normal	yes	no	5-12	<5	...	Asy/Patchy	TGA	no
1	1	no	Equal	Moderate	High	Grd_Glass	no	no	<5	5-12	...	Grd_Glass	Fallot	no
2	2	no	Equal	Severe	Normal	Plethoric	no	yes	5-12	5-12	...	Normal	PFC	no
3	3	no	Equal	Moderate	Normal	Plethoric	no	no	5-12	<5	...	Plethoric	PAIVS	no
4	4	no	Equal	Moderate	Normal	Plethoric	no	yes	12+	5-12	...	Plethoric	PAIVS	no
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
14995	14995	no	Equal	Moderate	Normal	Normal	no	no	5-12	5-12	...	Normal	PAIVS	no
14996	14996	no	Equal	Moderate	Normal	Plethoric	no	yes	<5	5-12	...	Plethoric	Fallot	no
14997	14997	no	Equal	Moderate	High	Normal	no	no	5-12	<5	...	Normal	Fallot	yes

The snippet of the dataset

The whole given data is present in the string. To understand the dataset first of all the columns should be understood.

1. Birth Asphyxia: Lack of oxygen during birth can lead to complications, including brain damage.
2. HypDistrib: Poor blood distribution can indicate inadequate blood flow to organs.
3. Hypoxia in O2: Inadequate oxygen supply can cause low oxygen levels in the blood.
4. CO2: Abnormal levels of CO2 in the blood can indicate respiratory or metabolic issues.
5. Chest X-ray: Medical imaging test used to visualize the chest's internal structures.
6. Grunting: Noise made during exhalation is a sign of respiratory distress in infants.
7. LVH Report: Provides details on the size and function of the heart's left ventricle.
8. Lower Body O2: Oxygen levels are measured to assess tissue oxygenation.
9. RUQO2: Oxygen levels measured in the right upper quadrant, indicating tissue oxygenation status.
10. CO2 Report: Provides information on carbon dioxide levels in the blood.
11. X-ray Report: Provides details on any abnormalities or conditions observed in the X-ray.
12. Disease: Refers to any deviation from normal health.
13. Grunting Report: Documents observations and findings related to grunting behavior in infants.
14. Age: The chronological age of the infant is an important factor in assessing growth, development, and medical management.
15. LVH: Condition involving the thickening of the heart's left ventricle wall.

16. Lung Parenchyma: Functional tissue of the lungs where gas exchange occurs.
All of this data is used to analyze the data and hence predict the sickness of the infants by using the model. Here is the analysis of the various columns in the dataset:



Purpose :

The purpose of this Project is to increase the accuracy of predicting the sickness of infant babies for early diagnosis.

Process:

First of all to understand the data dynamics I have used various plots using the library Matplotlib to plot the graphs. In Birth Asphyxia the number of infants suffering is more. The number of infants suffering from the disease Fallout is highest whereas those suffering from PFC are the least. Similarly in the X-ray report, the number of infants having Oliagiamic X-ray is more than the rest. These kinds of Data Analysis can be done using the graphs. Hence graphs play an essential role in the Data analytics field.

Later this data is encoded using comparable codes so that it will be suitable to train the the data and fit it properly in the algorithm. This coding is done by giving proper weights to the subgroups present in the data. The first model used to train the data is the Random Forest Classifier using the Sklearn library. Random forests are a favored supervised machine learning algorithm. Random forests are for supervised machine learning, where there is a labeled target variable. Random forests can be used for solving regression and classification problems. Random forests are an ensemble method, meaning they merge the predictions from other models. Each of the smaller models in the random forest ensemble is a decision tree. The other model used for training the data is the K-nearest neighbor algorithm. The K-Nearest Neighbors (KNN) algorithm is a popular machine-learning technique used for classification tasks. It relies on the idea that similar data points tend to have similar labels or values(the birds of the same feathers flock together). When making predictions, it calculates the distance between the input data point and all the training examples, using a chosen distance metric such as Euclidean distance.

To increase the accuracy of the model-built Feature Selection methods, Hyper Parameter Tuning is also used.

In Conclusion, the accuracy scores of the models are given below:
Random Forest Classifier: 0.7256666666666667
K-Nearest Neighbor: 0.7676666666666667
After Feature Selection: 0.7628
After Hyperparameter Tuning: 0.7693333333333333

From the accuracy scores it is very visible that the accuracy is increased after applying the suitable algorithms. As predicted, the accuracy should lie between 0.7-0.9 is achieved using the methods.