ISSN No:-2456-2165

# Prediction of Infant Death in Incubator using Machine Learning Techniques

O. D. Salunkhe[1]; P. H. Mahadik[2]
Department of Statistics, Yashavantrao Chavan Institute of Science
Satara- 415001

**Abstract:- Deaths in incubator is a major problem worldwide which is the leading cause of death in infants. Early diagnosis of infant death in an incubator helps to improve the survival chances of neonatal babies in an infant incubator. Early diagnosis of death in an incubator can save the infant or treat him/her better than standard treatment. In this paper we proposed a different ML algorithm to predict infant death in an incubator. This research was carried out in the Satara district, Maharashtra. In this study, we have taken factors that were identified by medical professionals for infant incubation. The machine learning algorithms can help medical professionals to predict infant deaths in an incubator as well as identify factors that cause the death of the infant in the incubator. The various data imbalancing techniques, such as the synthetic oversampling technique (SMOTE) and adaptive synthetic (ADYSN) have been implemented to improve the performance of models. The XG Boost classifier (F1 score = 0.88) and Random Forest classifier (F1 score = 0.89) with ADYSN give us the better performance than other classifiers.**

***Keywords:-*** *Infant Deaths, Incubator, Machine Learning Algorithm, Imbalanced Data.*

## I. INTRODUCTION

To ensure that every child survives and thrives and reaches their full potential, we need to focus on enhancing care during birth and the first week of life. Newborn deaths account for 47% of deaths among children under the age of five worldwide, costing 2.4 million lives each year. Approximately one-third of newborn deaths occur on the day of birth, with nearly three-quarters occurring within the first week of life. In addition, nearly 2 million babies are born with no signs of life at 28 weeks or more of pregnancy (stillbirths), and 295 000 mothers die each year [1].

Children who die within the first 28 days of birth suffer from conditions and diseases associated with lack of quality care at birth or skilled care and treatment immediately after birth and in the first days of life. The vast majority of newborn deaths take place in low and middle-income countries.

Machine learning is a computer science fiction; in real life, it is an area of study with applications in every field. Machine learning techniques are being used in a variety of applications, including signal processing, image and language recognition, industrial automation, and self-driving vehicles. The aim of this paper is to discover the method that provides the most accurate information about infant deaths in incubator. The concept behind this study is that if we can predict infant deaths as early as possible, we can reduce the risk and begin treatment as soon as possible. Using machine learning techniques, we can also reduce the time it takes to diagnose and handle massive amounts of medical data with ease.

Python has been chosen to implement the project due to it is easy to learn, understand, and implement. Python is open source and free, and it includes a number of machine learning libraries. In order to predict infant deaths in incubator, we trained the model on the dataset that was used for training using different techniques such as Logistic Regression (LR), K Nearest Neighbours (KNN), Decision Trees, SVM classifier, XG boost classifier and Random Forest and we tested the model's accuracy on the testing data set.

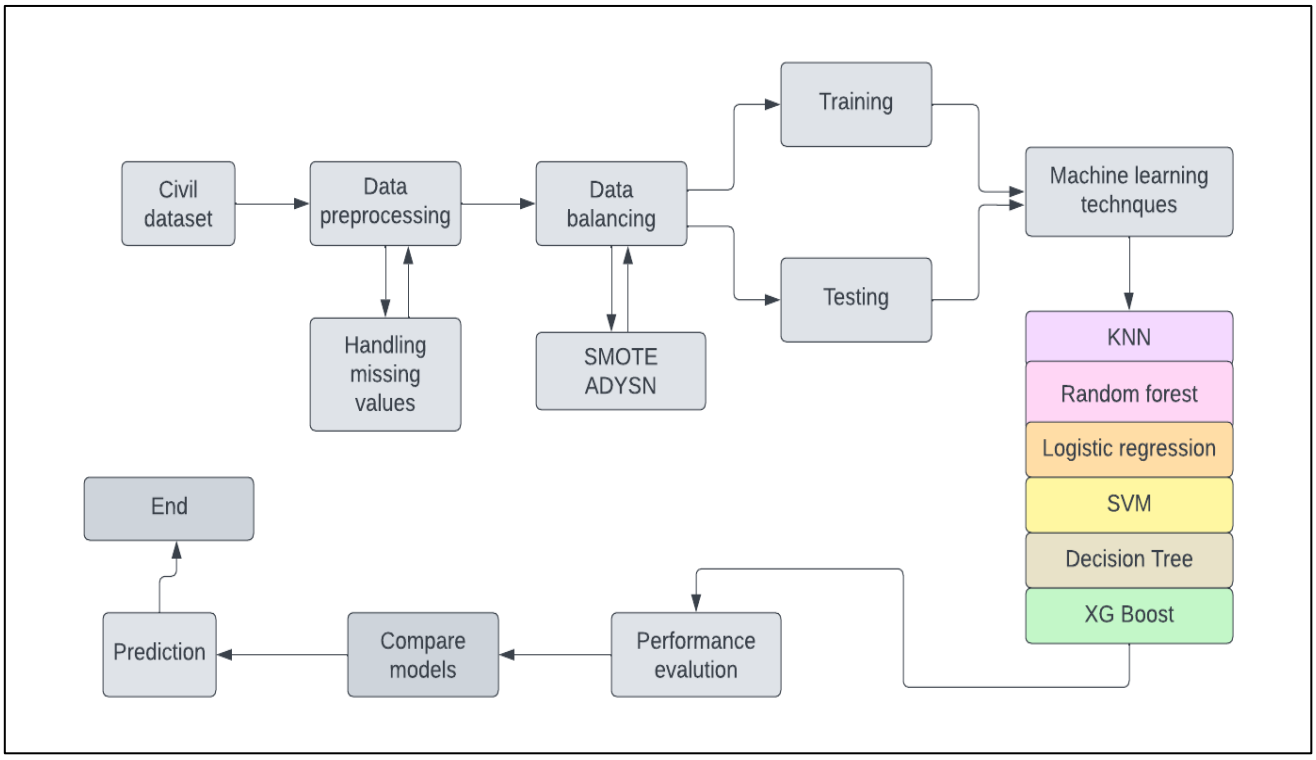


**Fig 1. Flowchart of Machine Learning Techniques**

## II. DATA COLLECTION

In this dataset containing the medical records of 1503 infants in incubator it collected from NICU department at the Kranti Singh Nana Patil Civil Hospital, Satara, Maharashtra from April 2022 to July 2023. The infants consist of 674 females and 839 males. The dataset contains 20 features, which report clinical, body, and lifestyle information, that we briefly describe here.

**Table 1: Data Description**

| Feature | Explanation |
|---|---|
| Gender | Male and female |
| Block | Rural and Urban |
| Type of admission | Inborn and Outborn |
| Maturity | Preterm: delivery before 37 completed weeks of gestation.<br>Full-term: Babies born between 37 week -42 weeks of gestation<br>Post-term: Any baby born after 42 weeks gestation. |
| Gestational age | Babies born in which week. |
| Age | Age of babies in days. |
| Weight | Weight of baby at birth. |
| Temperature | Temperature of baby at birth. |
| Heart rate | Heart rate of baby per minute at birth |
| RR | Respiratory rate per minute during birth. |
| Gravida | Number of times the woman has been pregnant. |
| Para | Indicates the number of births (including live births and stillbirths). |
| Livebirth | He completes expulsion or extraction from its mother |
| TT dose | Number of Vaccine for pregnant women. |
| HB | Haemoglobin during pregnancy |
| Labour | No: Normal labour usually begins within 2 weeks (before or after) the estimated delivery date<br>Induced: Inducing labour is prompting the uterus to contract during pregnancy before labour begins on its own for a vaginal birth.<br>Spontaneous: Vaginal delivery that happens on its own and without labour-inducing drugs |
| Delivery attended by | Doctor and Nurse |
| Time | Time of baby in incubator |
| Status | Alive and Death |

*A. Data Pre-Processing:*

The different models are developed on the original data but in the outcome variables the dead class having low frequency as compared to live class, the model performance is not good in F1 Score and precision. Hence, we used the data preprocessing techniques for best model.

The data is used to improve the model performance and assurance of machine learning models. For filling missing value or smoothing the data we used the mean, median for the quantitative data and mode is used for quantitative data. The missing values are filled in MS-Excel office.

*B. Data Balancing:*

Our datasets contain the frequency of alive cases is 1269 and the dead cases contain the frequency of 234. From the percentage of dead cases (15.56%) the data is unbalanced for the modelling. We used the different over sampling techniques such as Synthetic Over Sampling technique (SMOTE) and Adaptive Synthetic (ADYSN) to balance the data by using python version 3.11.

Based on above methods we created datasets in python. The original data set contains alive cases:1269, dead cases: 273 proportion:4.64, SMOTE datasets contain alive cases: 887, dead cases:887 proportion:1, ADYSN datasets contains alive cases:887, dead cases:921 proportion:0.96.

To achieve the best results, we used different supervised learning algorithms on original data sets and also used on SMOTE and ADYSN datasets and compared the accuracy, F1 score, precision and AUC scores.

*C. SMOTE: (Synthetic Over Sampling Technique)*

The majority class is under-sampled by randomly removing samples from the majority class population until the minority class becomes some specified percentage of the majority class. This forces the learner to experience varying degrees of under-sampling and at higher degrees of under-sampling the minority class has a larger presence in the training set [7].

*D. ADYSN: (Adaptive Synthetic)*

ADASYN is an extension of SMOTE tailored for imbalanced datasets. It generates more synthetic samples in areas with fewer instances, helping to balance class distribution effectively. ADASYN aims to improve model performance near decision boundaries. ADASYN has been shown to outperform traditional methods in handling class imbalance by generating synthetic samples intelligently. Its adaptive nature makes it particularly useful in real-world applications where class distributions may change over time.

*E. Model Development:*

In this step we used input variables to develop ML models. The data is divided into training and testing groups. In training group contains 70% and testing contains 30% and then applied various machine learning algorithms such as Logistic regression (LR), K nearest neighbour classifier (KNN), Random Forest classifier (RF), Decision tree classifier (DT), Support vector machine classifier (SVM), Gradient boost classifier (XG Boost).

*F. Logistic Regression:*

Logistic regression is a statistical method used for binary or dichotomous classification. Logistic regression models the probability of a binary outcome by applying the logistic function, also known as the sigmoid function. This function maps input values to probabilities between 0 and 1, making it suitable for binary classification problems.

During training, the model minimizes a loss function like binary cross-entropy. Once trained, it can predict the probability of the positive class for new instances. A decision threshold can then be chosen to convert probabilities into binary predictions.

*G. KNN:*

K-Nearest Neighbours (KNN) is a simple algorithm for classification and regression in machine learning. The basic idea behind KNN is to predict the class or value of a new data point by looking at the 'k' closest data points in the feature space. In classification tasks, the algorithm assigns the new data point to the most common class among its 'k' nearest neighbours. The effectiveness of KNN is determined by several distance metrics, including Euclidean, Manhattan, and Minkowski distances [8].

*H. Random Forest:*

Random Forest Classifier is a powerful ensemble learning method acting wide used for classification tasks in simple machine learning. It operates by constructing a throng of decision trees during grooming and outputs the mode of the classes (classification) of the person trees. Each decision tree in the forest is skilled on a random subset of the training information and a random subset of the features [5]. This randomness introduces diversity among the trees, reduction the risk of over fitting and improving stimulus generalization performance. Additionally, the decision trees are typically constructed using a subset of features at each node, boost enhancing the diversity and robustness of the model. Moreover, Random afforest Classifier is relatively insensitive to hyper parameter tuning, making it easier to use compared to other complex models.

*I. Decision Tree:*

It classifies data into tree like structure algorithm. Its workings by partitioning the feature space into segments supported on the values of input features, creating a tree-like structure where each internal node represents a decision based on a feature, and each leaf node represents the assort tag or predicted value. Decision trees are spontaneous and easy to interpret, making them valuable for understanding the underlying patterns in the data.

*J. SVM:*

Support Vector Machine (SVM) classifier is a powerful supervised learning algorithm used for both classification and regression tasks. It works by finding the hyperplane that best separates the data into different classes while maximizing the margin, which is the distance between the

hyperplane and the nearest data points (support vectors). SVM is more accurate for unbalanced dataset. In this we used linear kernel function and in tunning method we used radial basis function (RBF). SVMs are especially effective in high-dimensional spaces, where they can capture complex relationships in data by transforming input features into a higher-dimensional space using various kernel functions. Common kernel functions include linear and sigmoid.

*K. XG Boost:*

XG Boost (extreme Gradient Boosting) is an extremely effective simple machine algorithm known for its speed and performance. It is a boosting-based ensemble method acting in which weak learners are sequentially combined to improve predictive accuracy. To prevent overfitting, it uses regularization techniques much as L1 and L2 regularization. The algorithm allows for duplicate processing and tree pruning, which results in faster training and less retentiveness usage.

## III. IMPLEMENTATION OF DATA ANALYSIS

*A. Accuracy*

Accuracy is the proportion of true events to the total number of cases tested. In this study, it was used to calculate model efficacy and confusion matrix measurements.

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

*B. Precision*

$$Precision = \frac{TP}{TP + FP}$$

*C. Recall*

$$Recall = \frac{TP}{TP + FN}$$

*D. F1 Score*

The F1 score corresponds to the inverse relationship between accuracy and recall. When dealing with unbalanced data, a higher F1 score predicts a better model. The harmonic average of recall and accuracy is calculated as;

$$F1\ Score = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

## IV. RESULT

Based on Doctor's and Paediatrics opinions, 21 important risk factors were identified and four of them were excluded from the analysis due to high missing data in the dataset.

25 significant risk factors were identified based on the judgments of doctors and paediatricians; five of them were eliminated from the study because the dataset had a high percentage of missing data.

*A. Description of Data*

**Table 2: Distribution of Quantitative Features in Dataset**

| Variable | Mean | S.D. |
|---|---|---|
| Gestational Age | 35.61 | 2.77 |
| Age | 2.02 | 3.99 |
| Weight | 2.27 | 0.64 |
| Temperature | 37.04 | 0.4 |
| Heart Rate | 143.93 | 13.96 |
| RR | 49.67 | 11.93 |
| Gravida | 1.63 | 1.01 |
| Para | 1.34 | 0.88 |
| Live Births | 1.13 | 0.87 |
| Abortion | 0.013 | 0.13 |
| TT | 1.99 | 0.025 |
| HB | 10.39 | 1.4 |
| Time | 7.83 | 7.87 |

Total 1502 babies are in the study in which with an average gestational age of 35.61±02.77 weeks. Average age of babies is 2.02±3.99 days, average weight is 2.27±0.64 kg, average temperature of babies is 37.04±0.4, average heart rate of babies is 143.93±13.96 per minute, average respiration rate of babies are 49.67±11.93, average haemoglobin level of mother 55.4% infants are male (52.1% are dead and 56.0% are alive respectively).83.0% infants are in incubator are from rural area (84.2% are dead and 82.7% are alive).70.5% infants are in the incubator are inborn which are born in the hospital (68.4% are dead and 70.9% are alive). 68.5% infants in the incubator are preterm in which the 79.5% are dead and 65.9% are alive. 91.2% deliveries are attended by doctors but 89.5% are dead and 91.5% are survived.

ISSN No:-2456-2165

**Table 3**: **Distribution of Qualitative Features in Dataset**

| Variables | Values | Dead | | Alive | | Total | |
|---|---|---|---|---|---|---|---|
| | | Frequency | % | Frequency | % | Frequency | % |
| Gender | Female | 112 | 47.9 | 558 | 44 | 670 | 44.6 |
| | Male | 122 | 52.1 | 710 | 56 | 832 | 55.4 |
| Block | Rural | 197 | 84.2 | 1049 | 82.7 | 1246 | 83 |
| | Urban | 37 | 15.8 | 219 | 17.3 | 256 | 17 |
| Type of admission | Inborn | 160 | 68.4 | 899 | 70.9 | 1059 | 70.5 |
| | Outborn | 74 | 31.6 | 369 | 29.1 | 443 | 29.5 |
| Maturity | Full term | 48 | 20.5 | 428 | 33.8 | 476 | 31.7 |
| | Postterm | 0 | 0 | 4 | 0.3 | 4 | 0.3 |
| | Preterm | 186 | 79.5 | 836 | 65.9 | 1022 | 68 |
| Labour | Induced | 14 | 6 | 102 | 8 | 116 | 7.7 |
| | No | 105 | 44.9 | 512 | 40.4 | 617 | 41.1 |
| | Spontaneous | 115 | 49.1 | 654 | 51.6 | 769 | 51.2 |
| Delivery attended by | Doctor | 210 | 89.7 | 1160 | 91.5 | 1370 | 91.2 |
| | Nurse | 24 | 10.3 | 108 | 8.5 | 132 | 8.8 |

*B. Performance and Evaluation:*

The performance of selected models from original data are not good due to imbalanced data. Therefore, we created a two-dataset considering SMOTE and ADYSN datasets and validate the performance the performance of original data, SMOTE data and ADYSN data are compared in given table.

From maximum models developed on SMOTE and ADYSN show better than other datasets. In this the random forest classifier and XG boost classifier show best performance among following models, The XG boost classifier shows 0.88 accuracy and 0.88 F1 score in ADYSN data. The random forest classifier shows 0.88 accuracy and 0.89 F1 score in ADYSN data also 0.88 and 0.88 accuracy and F1 score is in SMOTE data.

**Table 4: Performance Measures of Machine Learning Algorithms**

| Classifier | Technique | Accuracy | Precision | F1 score | AUC score |
|---|---|---|---|---|---|
| Decision Tree | Original data | 0.76 | 0.28 | 0.31 | 0.59 |
| | Tunning | 0.84 | 0.5 | 0.25 | 0.60 |
| | SMOTE | 0.75 | 0.78 | 0.72 | 0.81 |
| | ADYSN | 0.74 | 0.74 | 0.75 | 0.74 |
| KNN | Original data | 0.83 | 0.33 | 0.10 | 0.60 |
| | Tunning | 0.84 | 0.41 | 0.17 | 0.60 |
| | SMOTE | 0.81 | 0.75 | 0.83 | 0.91 |
| | ADYSN | 0.80 | 0.74 | 0.83 | 0.88 |
| Random Forest | Original data | 0.86 | 0.57 | 0.28 | 0.70 |
| | Tunning | 0.85 | 0.23 | 0.33 | 0.62 |
| | SMOTE | 0.88 | 0.87 | 0.88 | 0.95 |
| | ADYSN | 0.88 | 0.88 | 0.89 | 0.94 |
| Logistic Regression | Original data | 0.87 | 0.66 | 0.29 | 0.60 |
| | Tunning | 0.84 | 0.4 | 0.79 | 0.70 |
| | SMOTE | 0.75 | 0.76 | 0.74 | 0.80 |
| | ADYSN | 0.72 | 0.73 | 0.73 | 0.79 |
| SVM | Original data | 0.87 | 0.66 | 0.29 | 0.60 |
| | Tunning | 0.84 | 0.50 | 0.78 | 0.67 |
| | SMOTE | 0.54 | 0.56 | 0.51 | 0.70 |
| | ADYSN | 0.56 | 0.58 | 0.56 | 0.59 |
| XG boost | Original data | 0.85 | 0.56 | 0.83 | 0.65 |
| | Tunning | 0.85 | 0.58 | 0.31 | 0.64 |
| | SMOTE | 0.87 | 0.86 | 0.87 | 0.95 |
| | ADYSN | 0.88 | 0.88 | 0.88 | 0.94 |

By using feature ranking of random forest classifier, we get some important variables for the classification.

ISSN No:-2456-2165

**Table 5: Importance of Each Feature**

| Feature | Information Gain |
|---|---|
| Weight | 0.1246 |
| Time | 0.1237 |
| HB | 0.0980 |
| Gestational Age | 0.0921 |
| RR | 0.0919 |
| Heart Rate | 0.0863 |
| Age | 0.0543 |
| Labour | 0.0534 |
| Sex | 0.0452 |
| Gravida | 0.0428 |
| Live Births | 0.0356 |
| Para | 0.0341 |
| Temperature | 0.0303 |
| Block | 0.0285 |
| Type of Admission | 0.0256 |
| Maturity | 0.0245 |
| Delivery Attended By | 0.0081 |
| Abortion | 0.0008 |

By using feature ranking weight, time, HB, Gestational age are mostly significant feature which are affected on infant deaths by using random forest classifier in ADYSN dataset

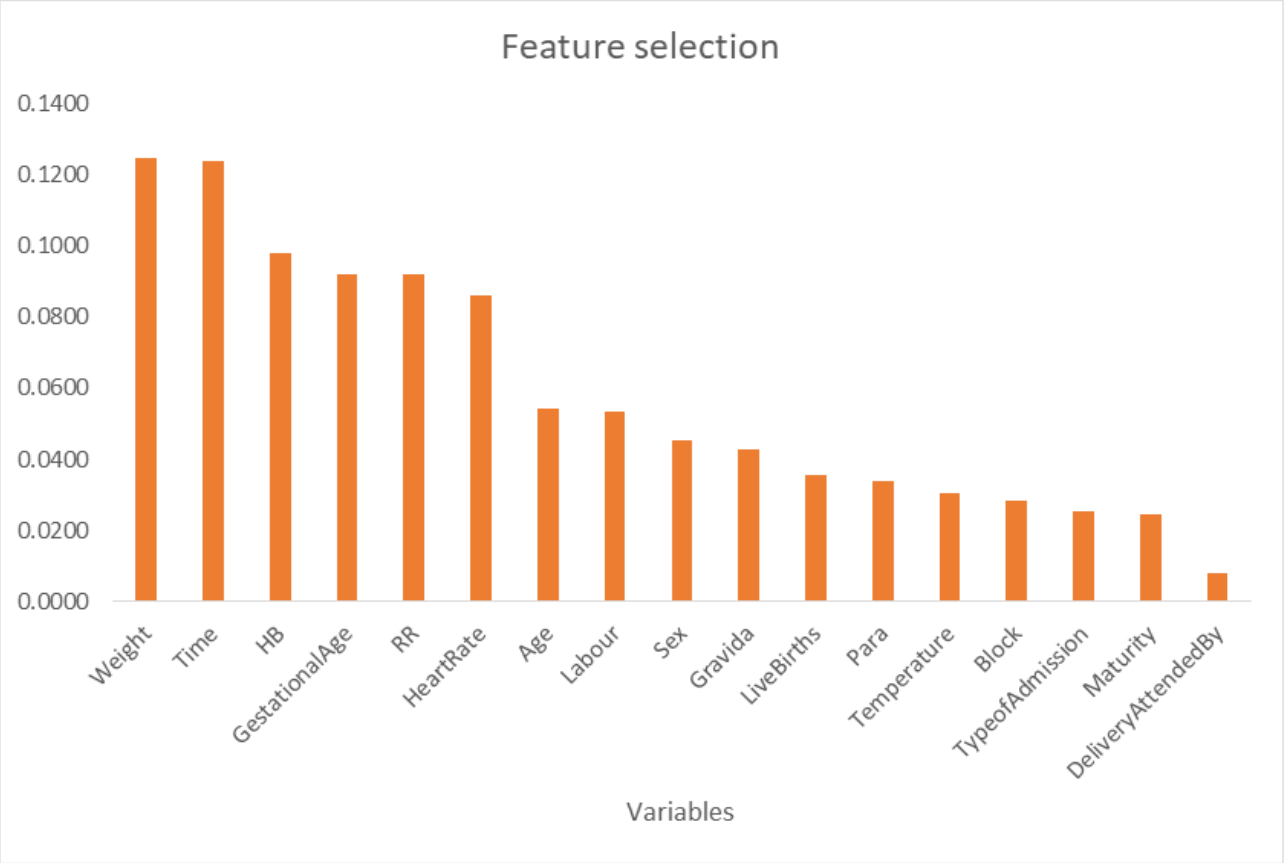


**Fig 2: Importance of Variables by Final Classification**

## V. CONCLUSION

In this study we used various supervised machine learning models like KNN, Decision tree classifier, Random Forest classifier, Logistic regression, XG Boost classifier, Support vector machine classifier using this to predict neonatal deaths in incubator. Models was developed using variables that important by doctors.

## REFERENCES

[1]. https://www.who.int/health-topics/newborn-health#tab=tab_1
[2]. Horbar JD, Edwards EM, Greenberg LT, et al. Variation in Performance of Neonatal Intensive Care Units in the United States. JAMA Pediatr. 2017;171(3):e164396. doi:10.1001/jamapediatrics.2016.4396
[3]. Haibo He, Yang Bai, E. A. Garcia and Shutao Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, 2008, pp. 1322-1328, doi: 10.1109/IJCNN.2008.4633969.
[4]. H. A. Gameng, B. B. Gerardo and R. P. Medina, "Modified Adaptive Synthetic SMOTE to Improve Classification Performance in Imbalanced Datasets," 2019 IEEE 6th International Conference on Engineering Technologies and Applied Sciences (ICETAS), Kuala Lumpur, Malaysia, 2019, pp. 1-5, doi: 10.1109/ICETAS48360.2019.9117287
[5]. Saroj, R.K., Yadav, P.K., Singh, R. *et al.* Machine Learning Algorithms for understanding the determinants of under-five Mortality. *BioData Mining* **15**, 20 (2022).
[6]. I. Hingorani, R. Khara, D. Pomendkar and N. Raul, "Police Complaint Management System using Blockchain Technology," 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), Thoothukudi, India, 2020, pp. 1214-1219, doi: 10.1109/ICISS49785.2020.9315884.
[7]. Chawla, Nitesh & Bowyer, Kevin & Hall, Lawrence & Kegelmeyer, W.. (2002). SMOTE: Synthetic Minority Over-sampling Technique. J. Artif. Intell. Res. (JAIR). 16. 321-357. 10.1613/jair.953.
[8]. https://medium.com/@nandiniverma78988/understanding-k-nearest-neighbors-knn-regression-in-machine-learning-c751a7cf516c