Fetal Health Classification Using Various Machine Learning Algorithm

Dr.P. Privietha[1], R. Monish[2], S. Mothish[3], R. Nikhil[4], M. Nithishkumar[5].

Dr.P. Privietha[1] Department of Computer Science and Engineering
Hindusthan College of Engineering and Technology
E-mail: priviethaprabhakar@gmail.com
R. Monish[2] Department of Computer Science and Engineering
Hindusthan College of Engineering and Technology
E. mail: 20104121@hicet.ac.in
S. Mothish[3] Department of Computer Science and Engineering
Hindusthan College of Engineering and Technology
E. mail: 20104122@hicet.ac.in
R. Nikhil[4] Department of Computer Science and Engineering
Hindusthan College of Engineering and Technology
E. mail: 20104124@hicet.ac.in
M. Nithishkumar[5] Department of Computer Science and Engineering
Hindusthan College of Engineering and Technology
E. mail: 20104127@hicet.ac.in

Abstract - ***The aim of this study is to compare various supervised machine learning algorithms for predicting the fetal health. In this study, the researcher utilizes numerical data to classify the health of the fetal using various supervised machine learning algorithms. By comparing the performance of different algorithms, such as KNearest Neighbors, Random Forest, Decision Tree and Logistic Regression, the best algorithm for classifying the fetal health. The study focuses on using numerical data such as baseline value, accelerations, fetal movement, uterine contraction, decelerations, short term variability to train and test the algorithms. The accuracy of each algorithm is measured, and the results are compared to identify the most effective algorithm for classification of fetal health. This study has the potential to provide valuable insights into the use of machine learning for prediction of fetal health and can assist healthcare professionals in making early and accurate diagnoses.***

Key Words*: *K-Nearest Neighbors, Random Forest, Logistic Regression, Decision Tree

1. INTRODUCTION

A fetal is basically an unborn offspring which is in the embryo stage until it comes to the world. During the pregnancy process, each three-month period is known by a name called trimester. During this process the fetus grows and develops and along with it the regular checkups are very important. As we all know that a pregnancy lasts for 9 months and in this long period there may be various reasons which may cause disability or mortality in the newborn which is a very severe case and this needs to be avoided

One of the main tools to analyze the health of the fetal in the womb is by doing a CTG (Cardiotocography) which generally is used to evaluate the heart beat and the uterine contractions hence the data generated is used by the doctor to analyze the health and give his wording. But there is a room for error hence the doctors are not reliable to analyze the data hence different machine learning algorithms have been there which can analyze the data and predict the fetal health based on it

This paper aims to predict and classify the fetal health using machine learning and classify the fetal health using machine learning algorithms. The study highlights the potential of machine learning in predicting fetal health and presents a comparative analysis of different methods to help healthcare professionals identify the most effective approach for early and accurate diagnosis

2. MACHINE LEARNING

Machine learning is a branch of artificial intelligence that focuses on creating algorithms and statistical models that allow computers to learn from data without being explicitly programmed. In other words, machine learning involves building mathematical models that enable computers to make predictions or decisions based on input data.

There are different types of Machine Learning techniques, including supervised learning, unsupervised learning, and reinforcement learning. Supervised learning involves training a model on a labeled dataset to predict the output for new input data. Unsupervised learning involves identifying patterns and relationships in an unlabeled dataset without a predetermined outcome. Reinforcement learning involves training a model to make decisions based on feedback received through interactions with an environment.

Machine learning has many applications in various fields, such as healthcare, finance, marketing, and transportation. In healthcare, machine learning is used to predict disease risk, diagnose illnesses, and develop personalized treatment plans. In finance, machine learning is used to detect fraud, predict stock prices, and assess credit risk. In marketing, machine learning is used to identify customer preferences and behavior to optimize campaigns. In transportation, machine learning is used to improve navigation systems and predict traffic patterns.

Overall, machine learning is a powerful tool that has the potential to revolutionize various industries by enabling more accurate predictions, decisions, and insights.

3. SUPERVISED ALGORITHM

Supervised learning is a type of machine learning technique that involves training a model on a labeled dataset to predict the output for new input data. In other words, the algorithm is "supervised" by providing it with a set of input output pairs (i.e., labeled data) and then training it to map the input to the correct output.

There are different types of supervised learning algorithms, including regression and classification. Regression algorithms are used to predict continuous values, such as the price of a house or the temperature. On the other hand, classification algorithms are used to predict discrete values, such as whether an email is spam or not, or whether a tumor is malignant or benign.

Supervised learning algorithms use various techniques, such as decision trees, linear regression, logistic regression, support vector machines, and neural networks. The performance of these algorithms is measured using metrics such as accuracy, precision, recall, and F1 score.

A. Logistic Regression

Logistic regression is a popular machine learning algorithm used for classification problems. In the case of fetal health classification, logistic regression is often used to predict whether a fetal health is likely to have normal based on certain features such as baseline value, accelerations, fetal movement, uterine contraction, decelerations.

The logistic regression model works by estimating the probability of the target variable (in this case, whether the fetal is normal or suspect or pathologic) based on the values of the predictor variables. The output of the model is a probability score between 0 and 1, which can be interpreted as the likelihood of the fetal health

The logistic regression algorithm uses a sigmoid function to transform the linear regression output into a probability score. The sigmoid function maps any real-valued number to a value between 0 and 1. If the probability score is greater than 0.5, the fetal health is classified as pathologic, and if it is less than 0.5, the fetal health is classified as normal.

Logistic regression is a useful algorithm for predicting the likelihood of a fetal health based on certain features. It can be trained on a dataset of fetal with known values and used to predict health of the fetal.

B. K Nearest Neighbors

K Nearest Neighbors (KNN) is a popular supervised machine learning algorithm used for classification tasks, including the classification of the fetal health. KNN algorithm makes predictions by finding the K number of closest training examples in the feature space and then assigning the class label based on the majority vote of their neighbors.

In the case of fetal health prediction, the algorithm can be trained on a dataset of patients with known data set values which is obtained from the CTG (Cardiotocography) report. The algorithm then uses this trained model to predict the likelihood of fetal health.

To use KNN algorithm for fetal health prediction, the first step is to collect and preprocess the data. The data can be sourced from various sources such as medical records or surveys. The data should be preprocessed by removing irrelevant features, dealing with missing data, and scaling the data to ensure that all features are on the same scale.

Once the data is preprocessed, the next step is to split the data into training and testing sets. The training set is used to train the KNN model, while the testing set is used to evaluate the performance of the model.

After the data is split, the KNN algorithm can be trained on the training set. The number of neighbors, K, is chosen, and the distance metric is defined, which can be Euclidean distance, Manhattan distance, or any other suitable metric.

Finally, the trained model can be used to predict the likelihood of fetal health based on data. The accuracy of the model can be evaluated by comparing the predicted values with the actual values of the testing set.

KNN algorithm is a useful supervised machine learning algorithm that can be used for the prediction of the fetal health. The algorithm requires the preprocessing of data, splitting the data into training and testing sets, defining the distance metric, and selecting the number of neighbors.

C. Decision Tree

Decision Tree is a type of supervised learning algorithm that is widely used in data mining and machine learning. It works by creating a tree-like model of decisions and their possible consequences. In the case of fetal health prediction, the algorithm uses various features such as baseline value, accelerations, fetal movement, uterine contraction, decelerations create a decision tree that predicts the health of the fetal.

The decision tree algorithm works by recursively splitting the data into smaller subsets based on the feature that best separates the data into classes. At each split, the algorithm selects the feature that has the highest information gain. The process continues until the data is separated into subsets that are as pure as possible.

Once the decision tree is constructed, it can be used to predict the class of new data by traversing the tree based on the values of the input features. The class that is associated with the leaf node reached by the traversal is then used as the predicted class.

Decision tree algorithm has been used for the fetal health prediction and has shown promising results. It is easy to interpret and understand, which makes it a popular choice for medical diagnosis and decision-making. However, it is prone to overfitting, which can be mitigated by using techniques such as pruning and cross-validation.

D. Random Forest

Random Forest is a supervised learning algorithm used for classification and regression tasks. It is a combination of multiple decision trees, where each tree is built using a random subset of features and data points. The algorithm works by creating a forest of decision trees and outputs the class that is the mode of the classes predicted by the individual trees.

In the case of fetal health prediction, Random Forest can be used to identify the important features that are highly correlated with the CTG (Cardiotocography) report of the fetal, such as baseline value, accelerations, fetal movement, uterine contraction, decelerations. These features can be used to train the model to predict whether the fetal health is normal or not

The Random Forest algorithm has several advantages, including its ability to handle missing data, outliers, and noisy data. It can also handle large datasets with high dimensionality, and can provide an estimate of feature importance. Additionally, the algorithm can be used for both classification and regression tasks.

Overall, Random Forest is a powerful algorithm that can be used for classification of the fetal health with high accuracy and can provide valuable insights.

4. LITERATURE REVIEW

Based on the data from the Kaggle.com that the computation of uterine contraction, and fetal heart characteristics on 2126 CTG saving, and classified by some professional obstetricians. Random Forest achieved the highest accuracy (around 95%), followed closely by KNN and Decision tree. Logistic Regression had the lowest accuracy among the algorithms.

An adaptive neuro-fuzzy inference system (ANFIS) demonstrated its performance by predicting normal and diseased states from CTG data with 97.2 percent and 96.6 percent accuracy, respectively. In it is saying that comparing the classifiers, random forest and XGBoost are performing well but the dataset used is imbalanced as some modified version is to be used to get the better output in terms of accuracy. In the dataset that has been is used is the CTG data which is observed to be beneficial to identify the abnormalities. The visual analysis along which the decision support system focuses has been made on the machine learning models that have been used.

5. DATASET

The dataset used in this study was obtained from Kaggle.com. This dataset contains 2126 records of features extracted from Cardiotocogram exams, which were then classified by expert obstetrician into 3 classes.

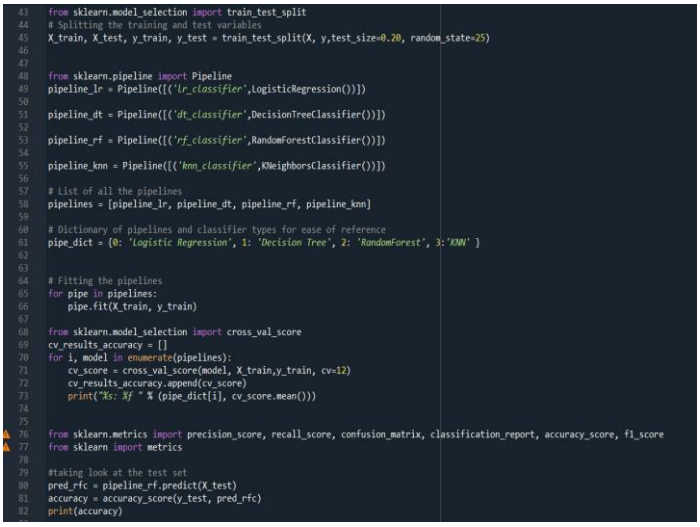
Cardiotocography (CTG) is used during pregnancy to monitor fetal heart rate and uterine contractions. It monitors fetal well-being and allows early detection of fetal distress.

On This Dataset Cardiotocograms (CTGs) are a simple and cost accessible option to assess fetal health, allowing healthcare professionals prevent child and maternal mortality. The equipment itself works by sending ultrasound pulses and reading its response, thus shedding light on fetal heart rate (FHR), fetal movements, uterine contractions and more.

6. METHODOLOGY

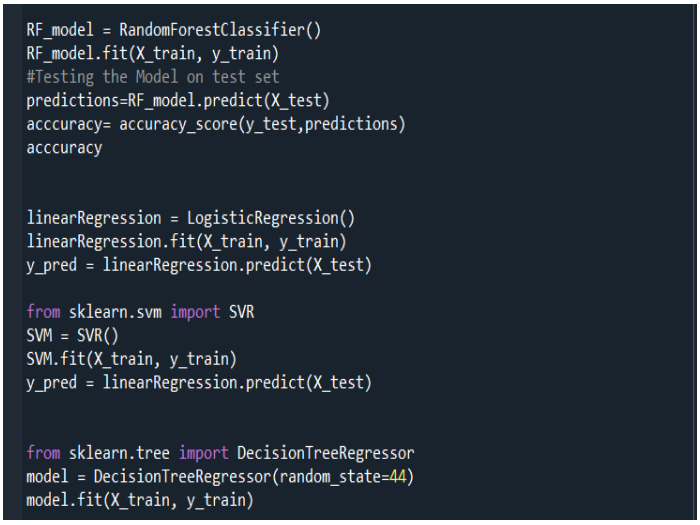
The experiment is carried out using python language in Spyder IDE. From anaconda tool the algorithm packages are imported. To conduct the analysis, the dataset was first collected from Kaggle.com in CSV format. The data was then split into two sets, with 80% of the data used for training and 20% for testing. Logistic regression was applied to the training data and the resulting model was evaluated based on its accuracy in classifying the test data.

Figure 1 represents the sample code in python language for training the dataset using logistic regression. The same process is followed for the other supervised algorithms like KNearest Neighbors, Random Forest, and Decision Tree. After training, testing stage is performed to calculate the accuracy.



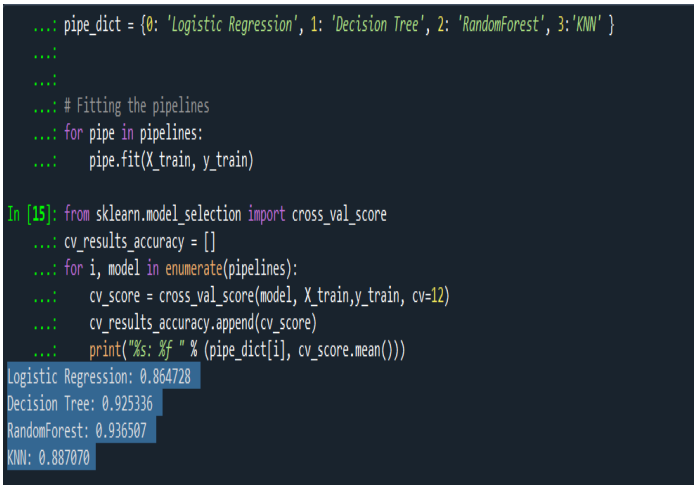
```
from sklearn.model_selection import train_test_split
# Splitting the training and test variables
X_train, X_test, y_train, y_test = train_test_split(X, y,test_size=0.20, random_state=25)


from sklearn.pipeline import Pipeline
pipeline_lr = Pipeline([('lr_classifier',LogisticRegression())])

pipeline_dt = Pipeline([('dt_classifier',DecisionTreeClassifier())])

pipeline_rf = Pipeline([('rf_classifier',RandomForestClassifier())])

pipeline_knn = Pipeline([('knn_classifier',KNeighborsClassifier())])

# List of all the pipelines
pipelines = [pipeline_lr, pipeline_dt, pipeline_rf, pipeline_knn]

# Dictionary of pipelines and classifier types for ease of reference
pipe_dict = {0: 'Logistic Regression', 1: 'Decision Tree', 2: 'RandomForest', 3:'KNN' }


# Fitting the pipelines
for pipe in pipelines:
    pipe.fit(X_train, y_train)

from sklearn.model_selection import cross_val_score
cv_results_accuracy = []
for i, model in enumerate(pipelines):
    cv_score = cross_val_score(model, X_train,y_train, cv=12)
    cv_results_accuracy.append(cv_score)
    print("%s: %f " % (pipe_dict[i], cv_score.mean()))


from sklearn.metrics import precision_score, recall_score, confusion_matrix, classification_report, accuracy_score, f1_score
from sklearn import metrics

#taking look at the test set
pred_rfc = pipeline_rf.predict(X_test)
accuracy = accuracy_score(y_test, pred_rfc)
print(accuracy)
```

Figure 2 represents the sample python code for testing the accuracy among the various machine learning algorithm



```
RF_model = RandomForestClassifier()
RF_model.fit(X_train, y_train)
#Testing the Model on test set
predictions=RF_model.predict(X_test)
acccuracy= accuracy_score(y_test,predictions)
acccuracy


linearRegression = LogisticRegression()
linearRegression.fit(X_train, y_train)
y_pred = linearRegression.predict(X_test)

from sklearn.svm import SVR
SVM = SVR()
SVM.fit(X_train, y_train)
y_pred = linearRegression.predict(X_test)


from sklearn.tree import DecisionTreeRegressor
model = DecisionTreeRegressor(random_state=44)
model.fit(X_train, y_train)
```

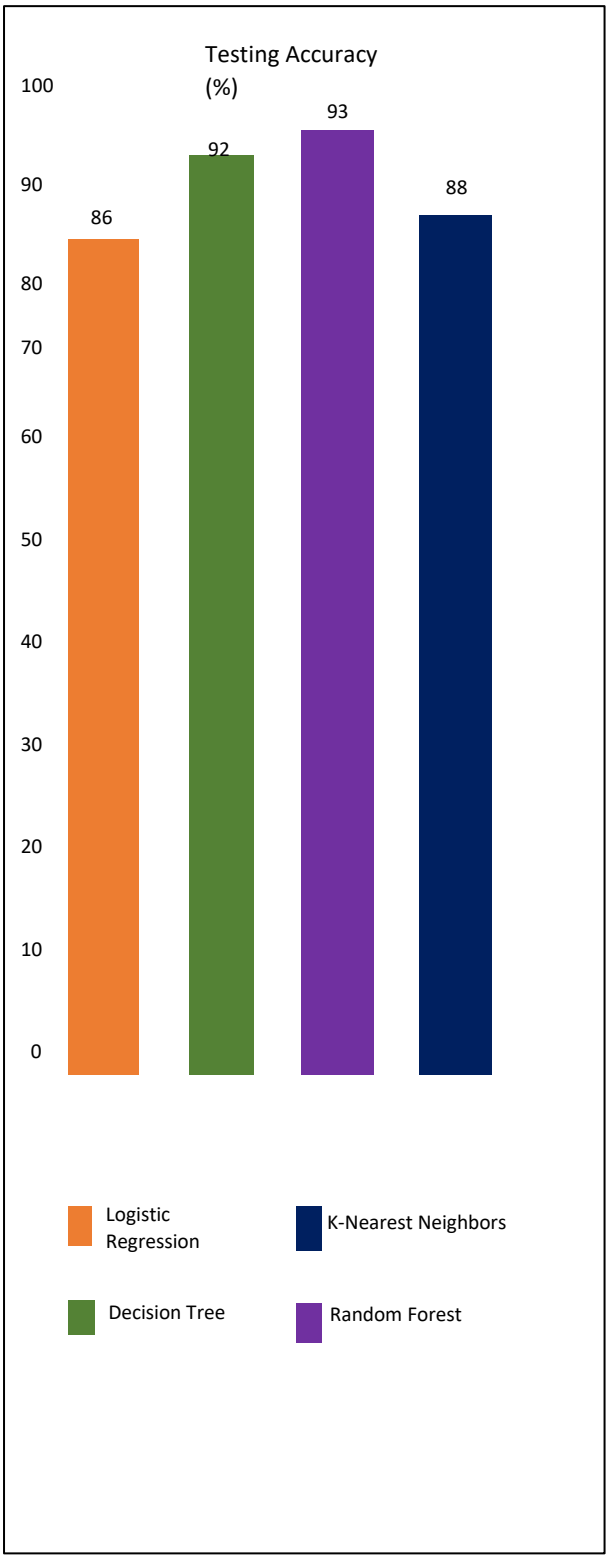
Figure 3 shows the testing accuracy of various algorithms like K-Nearest Neighbors, Random Forest, Decision Tree and Logistic Regression.



```
...: pipe_dict = {0: 'Logistic Regression', 1: 'Decision Tree', 2: 'RandomForest', 3:'KNN' }
...:
...:
...: # Fitting the pipelines
...: for pipe in pipelines:
...:     pipe.fit(X_train, y_train)

In [15]: from sklearn.model_selection import cross_val_score
...: cv_results_accuracy = []
...: for i, model in enumerate(pipelines):
...:     cv_score = cross_val_score(model, X_train,y_train, cv=12)
...:     cv_results_accuracy.append(cv_score)
...:     print("%s: %f " % (pipe_dict[i], cv_score.mean()))
Logistic Regression: 0.864728
Decision Tree: 0.925336
RandomForest: 0.936507
KNN: 0.887070
```

Figure 4 shows the graphical representation of testing accuracy of various algorithms like K-Nearest Neighbors,

Random Forest, Decision Tree and Logistic Regression. Among these algorithms Random Forest algorithm shows better accuracy of 93% while comparing other algorithm accuracy value like 86% for Logistic regression, 88% for K-Nearest Neighbors and 92% for Decision Tree Algorithm.



7.CONCLUSION

While comparing the accuracy of various algorithms like K-Nearest Neighbors, Random Forest, Decision Tree, and Logistic Regression. Among these algorithms Random Forest algorithm shows better accuracy of 93% in testing.

8. FUTURE WORK

There is potential for future research to expand upon the current work by incorporating an unsupervised dataset and addressing missing values by either deleting them or replacing them with valid data. Additionally, further optimization of accuracy through parameter selection could be explored while maintaining efficient training and testing times. Instead of relying solely on numerical features, deep neural networks and image recognition could be utilized for prediction. There is also the possibility of leveraging the developed machine learning models to create a licensed online portal for fetal health prediction and providing software to hospitals.

9.REFERENCE

[1] Privietha P, Joseph Raj V (2020), “Deep Learning Technic on Gait Analysis” published in Test Engineering and Management, Volume: 83, May –June 2020, SJR:0.1, ISSN: 0193-4120, pp: 11817 -11823.

[2] Privietha P, Joseph Raj V (2022), "Hybrid Activation Function in Deep Learning for Gait Analysis," 2022 International Virtual Conference on Power Engineering Computing and Control: Developments in Electric Vehicles and Energy Sector for Sustainable Future (PECCON), Chennai, India, 2022, pp. 1-7, doi: https://doi.org/10.1109/PECCON55017.2022.9851128.

[3] Mei-Ling Huang and Yung-Yan Hsu, Fetal distress prediction using discriminant analysis decision tree and artificial neural network, vol. 5, no. 9, pp. 526-533, 2012

[4] R Sindhu, J A Bahari, M Hariharan et al., "A Novel Clinical Decision Support System Using Improved Adaptive Genetic Algorithm for the Assessment of Fetal Well-Being", Computational and Mathematical Methods in Medicine, vol. 2015, pp. 1-11, 2015.

[5] M. Ramla, S. Sangeetha and S. Nickolas, "Fetal Health State Monitoring Using Decision Tree Classifier from Cardiotocography Measurements", 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 1799-1803, 2018..

[6] J. Piri, P. Mohapatra and R. Dey, "Fetal Health Status Classification Using MOGA - CD Based Feature Selection Approach", 2020 IEEE International Conference on Electronics Computing and Communication Technologies (CONECCT), pp. 1-6, 2020.